DOCUMENT RESUME

ED 406 454 TM 026 833

AUTHOR	Rogers, Alfred M.; And Others
TITLE	1990 Trial State Assessment Secondary-use Data Files User Guide. Revised June 1992.
INSTITUTION	Educational Testing Service, Princeton, N.J.; National Assessment of Educational Progress, Princeton, NJ.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
PUB DATE	Jun 92
NOTE	156p.
PUB TYPE	Guides - Non-Classroom Use (055) -- Statistical Data (110)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	*Data Analysis; Databases; Data Collection; *Educational Testing; Grade 8; Junior High Schools; Mathematics; National Surveys; Public Schools; Scaling; Scoring; State Programs; Tables (Data); *Test Construction; *Testing Programs; *User Needs (Information)
IDENTIFIERS	Data Files; *National Assessment of Educational Progress; Statistical Analysis System; Statistical Package for the Social Sciences; *Trial State Assessment (NAEP); User Guides

ABSTRACT

The National Assessment of Educational Progress (NAEP) is an ongoing, congressionally mandated national survey of the knowledge, skills, and understanding of young Americans in major subjects usually taught in school. This document is a guide to the files of the first NAEP Trial State Assessment (1990). The Trial State Assessment collected information on the mathematics knowledge, skills, understanding, and attitudes of a representative sample of about 100,000 students in public schools in 40 jurisdictions. The complexity of the assessment means that data file users need some special instructions. Following an overview of the NAEP system and the Trial State Assessment in the first chapter, the second chapter outlines special features of the assessment. The secondary-use data files for each state contain data for students, teachers, schools, and excluded students in the state and in the sample from the national mathematics assessment that was used for comparisons between the nation and the states. Other chapters discuss: (1) instrument design; (2) sample selection and weights; (3) data collection, processing, scoring, and database creation; (4) reporting subgroups and other variables; (5) NAEP scaling procedures; (6) conducting statistical analysis with NAEP data; (7) content and format of files, layouts, and codebooks; and (8) working with the Statistical Package for the Social Sciences and the Statistical Analysis System. Appendixes provide a history of the NAEP, the item response theory parameters for each cognitive item, and a glossary of terms. (Contains 29 tables and 22 references.) (SLD)

ED 406 454

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

NATIONAL CENTER FOR EDUCATION STATISTICS

National Assessment of Educational Progress

1990 Trial State Assessment Secondary-use Data Files User Guide

REVISED — JUNE 1992

Alfred M. Rogers
Debra L. Kline
John Mazzeo
Eugene G. Johnson
Robert J. Mislevy
Keith F. Rust

in collaboration with

John L. Barone
David S. Freund
Bruce A. Kaplan

BEST COPY AVAILABLE

TM 026833

Prepared by Educational Testing Service under contract with the National Center for Education Statistics
Office of Educational Research and Improvement • U.S. Department of Education

The work upon which this publication is based was performed for the National Center for Education Statistics, Office of Educational Research and Improvement, by Educational Testing Service.

Educational Testing Service is an equal opportunity/affirmative action employer.

Educational Testing Service, ETS, and ETS are registered trademarks of Educational Testing Service.

VERSION NOTICE

The June 1992 version of the NAEP 1990 Trial State Assessment secondary-use data files supersedes the version dated June 1991. Only minor revisions have been made for this version, none of which affect the original data values. The revisions are as follows:

- The Federal Information Processing Standards state identification code has been added to the student files (FIPS), the school files (SFIPS) and the excluded student files (XFIPS).

The content and format of the files are described in Chapter 9.

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
1990 TRIAL STATE ASSESSMENT
SECONDARY-USE DATA FILES USER GUIDE
Revised — June 1992

CONTENTS

List of Tables viii
Acknowledgments ix

Chapter 1 INTRODUCTION 3

1.1 What is NAEP? 3
1.2 Overview of the 1990 Trial State Assessment 3
  1.2.1 Special Features 4
1.3 The Secondary-use Data Files 5
1.4 Item Security 6
1.5 How to Use the Guide 6
1.6 An Analysis Example Using 1990 NAEP Data 8
  1.6.1 Beginning the Analysis 9
  1.6.2 Completing the Analysis with SAS 10
  1.6.3 Completing the Analysis with Other Statistical Languages 12
  1.6.4 Error Estimation 12

Chapter 2 SPECIAL CONSIDERATIONS FOR USERS 15

2.1 The National Comparison Sample of Students 15
2.2 Focused-BIB Spiral Method of Administration 15
2.3 Reporting Subgroups and Other Variables 15
2.4 Response Data from Teachers 16
2.5 Using Weights 16
2.6 Error Estimation 17
2.7 Monitored and Unmonitored Assessment Sessions 17

Chapter 3 INSTRUMENT DESIGN 21

3.1 Introduction 21
3.2 Student Assessment Booklets 21
  3.2.1 Booklet Content 21
  3.2.2 Booklet Assembly 22
  3.2.3 Release Status for Item Blocks 23
3.3 Questionnaires 23
  3.3.1 Student Questionnaires 23
  3.3.2 Excluded Student Questionnaire 25
  3.3.3 Teacher Questionnaire 26
  3.3.4 School Questionnaire 26

Chapter 4 SAMPLE SELECTION AND WEIGHTS 29

4.1 Introduction 29
4.2 Sample Selection 29
  4.2.1 Selection of Schools 30
  4.2.2 Selection of Student Samples 35
4.3 Weighting Procedures 36
  4.3.1 Full-sample Weights 36
  4.3.2 Comparison Weights for Monitored/Unmonitored Sessions 37
  4.3.3 Replicate Weights 38
  4.3.4 Summary of Weights and Their Use 39

Chapter 5 DATA COLLECTION, MATERIALS PROCESSING, PROFESSIONAL SCORING, AND DATABASE CREATION 43

5.1 Introduction 43
5.2 Data Collection and Field Administration 43
5.3 Materials Processing and Data Entry 43
5.4 Professional Scoring of Mathematics Items 44
  5.4.1 Description of Scoring 44
  5.4.2 Open-ended Scores in the Secondary-use Data Files 45
5.5 Database Creation 45

Chapter 6 REPORTING SUBGROUPS AND OTHER VARIABLES 49

6.1 Introduction 49
6.2 Reporting Subgroups for the 1990 Trial State Assessment 49
6.3 Variables Derived from the Questionnaires 53
6.4 Variables Derived from Mathematics Items 58
6.5 Variables Related to Proficiency Scaling 59
6.6 Principal's Questionnaire Variables 61

Chapter 7 NAEP SCALING PROCEDURES AND THEIR APPLICATION IN THE TRIAL STATE ASSESSMENT 65

7.1 Introduction 65
7.2 Theoretical Background of NAEP Scaling Procedures 65
7.3 Scaling Methodology 66
  7.3.1 The Scaling Model 67
  7.3.2 An Overview of Plausible Values Methodology 70
  7.3.3 Computing Plausible Values in IRT-based Scales 71
7.4 Analyses 73
  7.4.1 Computational Procedures 74
  7.4.2 Statistical Tests 75
  7.4.3 Biases in Secondary Analyses 76
7.5 Scale Anchoring 77
7.6 Scaling the 1990 Trial State Assessment Data 79
  7.6.1 Item Response Theory (IRT) Scaling 79

7.6.2 Estimation of State and Subgroup Proficiency Distributions 81
7.6.3 Linking State and National Scales 82
7.6.4 Producing a Mathematics Composite Scale 85
7.6.5 Proficiency Means for the 1990 Trial State Assessment Mathematics Scales 85

Chapter 8 CONDUCTING STATISTICAL ANALYSES OF 1990 NAEP TRIAL STATE ASSESSMENT DATA 89

8.1 Introduction 89
8.2 Using Weights to Account for Differential Representation 90
8.2.1 The 1990 State Samples of Students 91
8.2.2 Special Weights for Monitored and Unmonitored Sessions 92
8.2.3 The Winter Public-school Sample from the National Assessment 92
8.2.4 School-based Weights 93
8.3 Procedures Used by NAEP to Estimate Sampling Variability (Jackknifing) 93
8.3.1 Degrees of Freedom of the Jackknifed Variance Estimate 97
8.3.2 Estimation of Subpopulations with Appropriate Jackknifed Standard Errors 98
8.4 Procedures Used by NAEP to Handle Measurement Error 99
8.5 Approximations 102
8.6 A Note Concerning Multiple Comparisons 104

Chapter 9 CONTENT AND FORMAT OF DATA FILES, LAYOUTS, AND CODEBOOKS 107

9.1 Introduction 107
9.2 Data Files 107
9.2.1 Raw Data Files 109
9.2.2 SPSS-X and SAS Control Statement Files 109
9.2.3 Machine-readable Catalog Files 110
9.3 Printed Documentation 113
9.3.1 File Layouts 113
9.3.2 Codebooks 115

Chapter 10 WORKING WITH SPSS-X AND SAS 119

10.1 Introduction 119
10.2 SPSS-X and SAS Control Statement Files 119
10.3 Creating SPSS-X System Files 120
10.4 Creating SAS System Files 122
10.5 Merging Files Under SPSS-X or SAS 122
10.6 Computing the Estimated Variance of a Mean (Jackknifing) Using SPSS-X or SAS 124
10.7 An Analysis Example Using 1990 NAEP Data with SAS 131

Appendix A NAEP HISTORY 139
Appendix B 1990 TRIAL STATE ASSESSMENT IRT PARAMETERS 145
Appendix C GLOSSARY OF TERMS 155
References 163

List of Tables

Table 1-1 Participants in the 1990 Trial State Assessment Program 4
1-2 Analysis example 8
1-3 NAEP variables used to produce the analysis 10
1-4 SAS code for steps 3, 4, and 5 to produce sample analysis 11

3-1 Content-by-ability distribution of items, grade 8 22
3-2 Cognitive and noncognitive block information 24
3-3 Contents of assessment booklets 24

4-1 Summary of weights for the 1990 Trial State Assessment 39

6-1 NAEP regions 53
6-2 Scaling variables for the 1990 Trial State Assessment samples 60

7-1 Three example items for scale anchoring 78
7-2 Weights for the composite scale 85
7-3 Average mathematics proficiencies by scale and plausible value for the 1990 national winter public-school comparison sample 86

8-1 Example dataset to demonstrate the jackknife 96
8-2 Factors used to calculate the rescaled weight from the original weight 99

9-1 NAEP 1990 state data package description 108
9-2 Special response codes 109
9-3 NAEP 1990 state machine-readable catalog file layout 111
9-4 Scaling categories and codes 113

10-1 SPSS-X control statement synopsis 121
10-2 SAS control statement synopsis 123
10-3 Matching school and student files 125
10-4 Matching school and excluded student files 126
10-5 Standard error computation: Multiweight method using SPSS-X 127
10-6 Standard error computation: Multiweight method using SAS 128
10-7 Standard error computation: Multiweight method using SPSS-X with correction for imputation 129
10-8 Standard error computation: Multiweight method using SAS with correction for imputation 130
10-9 SAS analysis example using jackknifed standard error estimates 131
10-10 SAS code for steps 2 through 7 to produce sample analysis 134

ACKNOWLEDGMENTS

The NAEP 1990 Trial State Assessment data are a unique and valuable source of information for the research community. To extend the usefulness of these data, we have provided the secondary-use data files and user guide. We hope these data will be used by many researchers for secondary data analysis.

The data files were created through the efforts of a talented and dedicated team of data analysts. Special acknowledgment must be given to Alfred Rogers, who developed the sophisticated database systems and created the secondary-use data files, and to David Freund, who was responsible for the creation of the database for the state data. Special thanks go to Bruce Kaplan and Edward Kulick, who made significant contributions to the analysis of the data. Also providing data analysis support were Drew Bowker, Phillip Leung, and Wing Lowe.

The user guide has been particularly enhanced through the contributions of the Statistical and Psychometric Research staff members John Mazzeo, Eugene Johnson, and Robert Mislevy. We are also grateful to Stephen Koffler and Keith Rust for their contributions and comments.

The user guide was designed and produced under the outstanding editorial supervision of Debra Kline.

John L. Barone
Director of Data Analysis
National Assessment of
Educational Progress

Chapter 1

INTRODUCTION

Chapter 1: INTRODUCTION

1.1 WHAT IS NAEP?

The National Assessment of Educational Progress (NAEP) is an ongoing, congressionally mandated national survey of the knowledge, skills, understanding, and attitudes of young Americans in major subjects usually taught in school. Its primary goals are to detect and report the current status of and long-term changes in the educational attainments of young Americans. The purpose of NAEP is to gather information that will aid educators, legislators, and others in improving the educational experience of youth in the United States. It is the first ongoing effort to obtain comprehensive and dependable achievement data on a national basis in a uniform, scientific manner.

NAEP began in 1969 as an annual survey of American students ages 9, 13, and 17 in various subject areas; young adults ages 26 to 35 were surveyed less frequently. Since the 1980-81 school year, budget restraints have prompted a shift to biennial data collection. In the 1984 assessment, NAEP began sampling students by grade as well as age.

In April 1988, Congress reauthorized NAEP and added a new dimension to the program -- voluntary state-by-state assessments on a trial basis, in addition to continuing the national assessments that NAEP has conducted since its inception.

More information about NAEP and its history is provided in Appendix A.

1.2 OVERVIEW OF THE 1990 NAEP TRIAL STATE ASSESSMENT

The first NAEP Trial State Assessment was conducted between February 5 and March 2, 1990. The program collected information on the mathematics knowledge, skills, understanding, and attitudes of a representative sample of eighth-grade students in public schools in 40 jurisdictions[1] -- 37 states, the District of Columbia, and two territories (shown in Table 1-1). National assessments in mathematics, reading, writing, and science were conducted simultaneously at age 9/grade 4, age 13/grade 8, and age 17/grade 12.

The Trial State Assessment data were collected from more than 100,000 students across the 40 states, based on a complex sample survey. The students who were assessed were administered one of seven mathematics assessment booklets that were also used in NAEP's 1990 national mathematics assessment.

---

[1] The word *state* is used throughout this document to refer to any of the 40 jurisdictions that participated in the assessment, even though three of them -- the District of Columbia, Guam, and the Virgin Islands -- are not states.

Table 1-1
Participants in the
1990 Trial State Assessment Program

States and Other Jurisdictions			
Alabama	Guam	Minnesota	Oklahoma
Arizona	Hawaii	Montana	Oregon
Arkansas	Idaho	Nebraska	Pennsylvania
California	Illinois	New Hampshire	Rhode Island
Colorado	Indiana	New Jersey	Texas
Connecticut	Iowa	New Mexico	Virginia
Delaware	Kentucky	New York	Virgin Islands
District of Columbia	Louisiana	North Carolina	West Virginia
Florida	Maryland	North Dakota	Wisconsin
Georgia	Michigan	Ohio	Wyoming

The mathematics framework and objectives established to guide both the Trial State Assessment and national assessment were developed for NAEP through a consensus project of the Council of Chief State School Officers, funded by the National Center for Education Statistics and the National Science Foundation. The framework and objectives were also used for the 1990 national mathematics assessment. In addition, questionnaires completed by the students, their mathematics teachers, and principals or other school administrators provided an abundance of contextual data within which to interpret the mathematics results.

Educational Testing Service (ETS) was the contractor for the 1990 NAEP program, including the Trial State Assessment Program. ETS was responsible for overall management of the programs as well as for development of the overall design, the items and questionnaires, data analysis, and reporting. Westat, Inc., and National Computer Systems (NCS) were subcontractors to ETS. Westat was responsible for all aspects of sampling and of field operations, while NCS was responsible for printing, distribution, receipt, professional scoring of the open-ended items, and data entry of all assessment materials.

1.2.1 Special Features

Because of the complexity of the NAEP design (see Chapters 3 and 4), data file users need some understanding of the design before performing analyses. Special characteristics of the assessment are outlined in Chapter 2.

The data files contain sampling weights for each student that should be used in statistical analyses. In addition, because of the complex sampling scheme, conventional methods of standard error estimation do not produce appropriate estimates. The NAEP sampling design also reduces the effective degrees of freedom for statistical analysis. These issues are discussed in Chapter 8.

1.3 THE NAEP 1990 SECONDARY-USE DATA FILES

Historically, a "public-use" version of the NAEP data files has been distributed to secondary users. However, in order to comply with 5 U.S.C. 552a and U.S.C. 1221e-1, only a "restricted-use" version of the 1990 NAEP data files will be distributed for secondary use. These will be loaned to states and people designated by them under a new licensure procedure designed to assure confidentiality of identifiable district, school, and individual data.

The secondary-use data files for each state contain data for students, teachers, schools, and excluded students in the state and for students, teachers, and schools in the sample from the national mathematics assessment that was used for comparisons between the nation and the state. The June 1992 version of the files supersedes the previous version, dated June 1991. The secondary-use data files contain:

- students' responses to cognitive mathematics items;
- students' responses to questions about their demographic backgrounds and educational experiences;
- information about students' schools and mathematics teachers;
- information about students excluded from the assessment (state samples only);
- sampling weights for students, schools, and (for state samples only) excluded students;
- proficiency scale scores for the mathematics composite scale and each of the five mathematics subscales -- Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions;
- machine-readable catalog files; and
- SPSS-X and SAS control statement files.

The data files are available in compressed (PKZIP) format on high-density 3.5" or 5.25" diskettes and in blocked EBCDIC format on nine-track tape reels (recording density of 6,250 bytes per inch) or an IBM 3480 tape cartridge (recording density of 38,000 bytes per inch). To use the NAEP data files, you will need an IBM DOS or MS-DOS-type microcomputer with a 5.25" or 3.5" high-density drive or either a mini- or mainframe computer with the appropriate tape drive.

Codebooks for each state provide the layout of the data, a description of each variable, and a description of each raw data file for both the state and the sample from the national mathematics assessment that was used for comparisons between the nation and the state. The content and format of the data files and codebooks are described in Chapter 9. Table 9-1 in that chapter gives the number of files for each sample and the record lengths for each file.

If you have questions about the data files and their use, contact one of the following individuals:

Mr. John Mathews
Statistician, National Center for Education Statistics
(202) 219-1690

Mr. Kent Ashworth
Director of Dissemination, National Assessment of Educational Progress
(609) 734-1327

1.4 ITEM SECURITY

In accordance with federal legislation regarding security of NAEP items and guidelines designed by the National Center for Education Statistics, each NAEP cognitive item has been assigned a release status. *Public release* items are available for unrestricted public use. *Secured release* items are available only to users who have agreed to conditions designed to ensure item security and to prevent misuse of items. *Non-released* items are those reserved exclusively for NAEP use -- for example, for administration in future assessments to permit analysis of trends in performance levels. To preserve the integrity of NAEP, it is essential that these items remain secure.

The data files and codebooks contain response counts for all items used in the assessment and descriptive text (the item stem and distractors) for each public release or secured release item. For each cognitive mathematics item that has been classified as non-released, the item stem and distractors have been replaced with short descriptions.

All student demographic and mathematics background items and items from teacher, school, and excluded student questionnaires are classified as public release and are available to the secondary user.

1.5 HOW TO USE THE GUIDE

Chapters 2 through 10 and the appendices provide detailed information about the 1990 Trial State Assessment, the data files, and recommended methods of working with the data to perform analyses. A summary of these chapters follows.

Chapter 2: Special Considerations for Users

This chapter describes features of the assessment design and assessment data that may be of special concern to researchers who wish to perform their own analyses of the data.

Chapter 3: Instrument Design

This chapter includes a description of the content, organization, and method of administration for the student assessment booklets and the teacher, excluded student, and school questionnaires.

Chapter 4: Sample Selection and Weights

This chapter explains the methods by which schools, students and teachers were chosen to be included in the assessment; the method by which some students were chosen for the sample but subsequently excluded from the assessment; and the sampling weights included on the data files.

Chapter 5: Data Collection, Materials Processing, Professional Scoring, and Database Creation

Procedures used for administering assessments, methods used for data entry and editing, how open-ended response items were scored, and procedures used to create the NAEP database can be found in this chapter.

Chapter 6: Reporting Subgroups and Other Variables

This chapter describes the NAEP reporting subgroups, derived and composite variables from the background questionnaires, composite variables created for the NAEP reports, item response theory (IRT) variables, and other data variables that are not self-explanatory.

Chapter 7: NAEP Scaling Procedures and Their Application in the Trial State Assessment

This chapter provides an overview of the scaling methodologies used by NAEP, the scale-score analyses carried out in the 1990 Trial State Assessment, and supporting information on the scale-score variables that appear on the data files.

Chapter 8: Conducting Statistical Analyses with NAEP Data

This chapter discusses the weights on the data files, how to use them in different types of analyses, and methods for estimating sampling variability and measurement error.

Chapter 9: Content and Format of Data Files, Layouts, and Codebooks

Detailed descriptions of the raw data files, layouts, codebooks, machine-readable catalogs, and SPSS-X and SAS control statement files are found in this chapter.

Chapter 10: Working with SPSS-X and SAS

This chapter provides procedures for creating SPSS-X and SAS system files, merging files, and using the jackknife procedure to estimate standard errors, as well as an example of how to analyze NAEP data with SAS.

Appendix A provides information about the history of NAEP.

Appendix B contains IRT parameters for each cognitive item used in the scaling of the mathematics data.

Appendix C is a glossary of terms.

References provide complete information on sources cited in the text.

1.6 AN ANALYSIS EXAMPLE USING 1990 NAEP DATA

This section presents an example of how to produce a simple descriptive analysis table from the national winter public-school (NWP) data files that are used for state/nation comparisons. The example could be carried out in a similar way for each state's files. Most analyses of NAEP data can be performed in four basic steps:

- Identify and access the appropriate data file
- Identify and extract the relevant variables
- Select the proper subset of students
- Compute and print the results

The method you choose to perform these steps may vary with the complexity of the analysis or with the statistical or procedural language you are using.

To aid users, we have added three types of data files:

- machine-readable catalog files
- SAS control statement files
- SPSS-X control statement files

The machine-readable catalog files can be used with any statistical or procedural language to quickly extract and store the location and labeling information for every field on the NAEP data files. This information can then be used by your program to extract actual response data from the data files. There is a catalog file for each data file; each catalog file contains a record for every field in the corresponding data file.

For SAS and SPSS-X users, control statement files are provided to facilitate the creation of SAS and SPSS-X system files. There is a SAS and an SPSS-X control file for each data file.

Part of each control file contains the field name, location, and format for each variable on the corresponding data file (more about control statement files can be found in Chapter 10).

1.6.1 Beginning the Analysis

The analysis in our example produced an estimate of the mean mathematics proficiency level for eighth-grade public-school girls in the national winter public-school (NWP) sample by the amount of television watched each day (Table 1-2).

Table 1-2
Analysis Example

```
               1990 NATIONAL WINTER PUBLIC SCHOOL SAMPLE
                MATHEMATICS RESULTS FOR 8TH GRADE GIRLS
                    BY AMOUNT OF TELEVISION VIEWING

          HOW MUCH
          TELEVISION DO
          YOU USUALLY        WEIGHTED
   OBS    WATCH                 N         PERCENT      MEAN

    1     NONE                  8.652      0.6069    257.174
    2     1 HOUR OR LESS      188.371     13.2129    268.729
    3     2 HOURS             284.804     19.9770    266.664
    4     3 HOURS             334.463     23.4602    261.801
    5     4 HOURS             245.271     17.2040    259.166
    6     5 HOURS             153.439     10.7627    253.907
    7     6 HOURS OR MORE     210.659     14.7763    239.367
```

To begin this analysis, you need to identify

- the file that contains response data for the national comparison sample of eighth-grade public-school students and
- the relevant variables in the file.

NAEP files are described in Chapter 9 and listed in Table 9-1; the correct file for our example is 'NWPSTUD.DAT'. Next, find the data set record layout for 'NWPSTUD.DAT' in the accompanying codebook. Here you will find the names and file locations of the variables needed to produce this table (unweighted response counts for each variable are found in the corresponding codebook). Four variables (described in Table 1-3) are required to produce the table: DSEX, WEIGHT, B001801A, and MRPCMP1.

Table 1-3
NAEP Variables Used to Produce the Analysis

Seq. No.	Field Name	Column Position	Field Width	Decimal Places	Type	Range	Short Label
27	DSEX	56	1	–	D	1 - 2	GENDER
47	WEIGHT	110	7	5	C	–	OVERALL STUDENT FULL-SAMPLE WEIGHT
279	B001801A	1361	1	–	D	1 - 7	HOW MUCH TELEVISION DO YOU USUALLY WATCH EACH DAY
265	MRPCMP1	1327	5	2	C	–	PLAUSIBLE NAEP MATH VALUE #1 (COMPOSITE)

Because this example is relatively simple (requiring the use of only four variables), you can manually enter the variable labels and locations into your computer program. Analyses that require many variables are performed more efficiently through use of the machine-readable catalog files or, if you are a SAS or SPSS-X user, the control statement files.

Section 1.6.2 describes how to complete the analysis using the statistical package SAS. SPSS-X users can use SPSS-X procedures in a similar way to perform analyses. Section 1.6.3 describes how to use the machine-readable catalog files to complete the analysis using statistical or procedural languages other than SAS or SPSS-X. In section 1.6.4, we discuss the importance of the proper estimation of standard errors.

1.6.2 Completing the Analysis with SAS

You can use any statistical computing language or package to access the raw data file, extract the relevant variables, select the proper subset of students, and compute the table. In this section, we carry out the rest of the analysis using the statistical package SAS.

1) Select the file containing the national winter public-school sample students. This is one of the samples described in Table 9-1; its file name is 'NWPSTUD.DAT'. Identify the relevant variables from the data set record layout: DSEX, WEIGHT, B001801A, and MRPCMP1.

2) Using the raw data file 'NWPSTUD.DAT', select the appropriate subset of students for the table. This selection restricts the analysis to girls (DSEX=2) who have valid MRPCMP1 (mathematics proficiency) and B001801A (television viewing) values.

3) Using the SAS procedures FREQ and SUMMARY, produce weighted cell counts and mathematics proficiency means for each level of the variable B001801A.

4) The final procedure in the SAS program merges the two sets of statistics and prints the table in a concise, labeled format.

The code for performing steps 2 through 4 is shown in Table 1-4.

Table 1-4

SAS Code for Steps 2, 3, and 4 to Produce Sample Analysis

```
TITLE1 '1990 NATIONAL WINTER PUBLIC SCHOOL SAMPLE';
TITLE2 'MATHEMATICS RESULTS FOR 8TH GRADE GIRLS';
TITLE3 'BY AMOUNT OF TELEVISION VIEWING';
DATA    A;
INFILE  RAWDATA;
INPUT
  DSEX          56        WEIGHT      110-116 5
  B001801A   1361        MRPCMP1    1327-1331 2 ;
IF (MRPCMP1 NE .);
IF (DSEX  EQ 2);
IF (B001801A NE .) AND
   (B001801A GT 0) AND
   (B001801A LT 8);
KEEP DSEX WEIGHT B001801A MRPCMP1;
LABEL
  DSEX     = 'GENDER                                  '
  WEIGHT   = 'OVERALL STUDENT FULL-SAMPLE WEIGHT      '
  B001801A = 'HOW MUCH TELEVISION DO YOU USUALLY WATCH'
  MRPCMP1  = 'PLAUSIBLE NAEP MATH VALUE #1 (COMPOSITE)';
PROC FORMAT;
  VALUE DSEX      1='MALE                '    2='FEMALE              ';
  VALUE B001801A  .='TOTAL               '    1='NONE                '
                  2='1 HOUR OR LESS      '    3='2 HOURS             '
                  4='3 HOURS             '    5='4 HOURS             '
                  6='5 HOURS             '    7='6 HOURS OR MORE     ';
PROC FREQ;
    TABLES B001801A / OUT=B;
    WEIGHT WEIGHT;
PROC SUMMARY DATA=A;
    CLASS B001801A;
    VAR MRPCMP1;
    OUTPUT OUT=C
        MEAN(MRPCMP1)=XBAR;
DATA D;
    MERGE B C;
    BY B001801A;
    IF (B001801A NE .);
PROC PRINT SPLIT='*';
    FORMAT B001801A B001801A.;
    LABEL COUNT   = 'WEIGHTED N'
          PERCENT = 'PERCENT'
          XBAR    = 'MEAN';
    VAR  B001801A COUNT PERCENT XBAR;
```

Please note that this example does not include standard error estimates that account for NAEP sampling design and measurement error components. In Chapter 10, we provide a second version of this example that demonstrates the proper computation of standard error estimates.

1.6.3 Completing the Analysis with Statistical or Procedural Languages Other than SAS or SPSS-X

This section explains how to complete the sample analysis using the machine-readable catalog files. Each catalog file contains one record for every data field in its corresponding data file. These records describe the contents of each data field (e.g, field name, field location, response labels, range of data in the field, etc.). Table 9-3 contains a complete layout for the catalog files.

In our example, 'NWPSTUD.CAT' (see Table 9-1) is the machine-readable catalog file that corresponds to the student data file 'NWPSTUD.DAT'. Each of the records in this catalog file describes one of the fields in the student data file. To access the student data with the catalog file and complete the analysis:

1) Extract and store the field locations and labels for every field contained on the student data file by reading the entire catalog file into your program.

2) Using the stored information from the catalog file, read the student data file to extract and label all of the student data fields.

3) In your program, select the data fields you want to work with and perform the required analyses (DSEX, WEIGHT, B001801A, and MRPCMP1).

4) Print the results using the stored labeling information from the catalog file.

Please note that this procedure does not include standard error estimates that account for NAEP sampling design and measurement error components (see section 1.6.4).

1.6.4 Error Estimation

The preceding example is presented as a practical introduction to the secondary-use data files. We have not attempted here to produce proper standard error estimates that account for NAEP sampling design and measurement error components. Such an accounting is required for statistical comparison of the results shown in our table. Because the NAEP sample is not a simple random sample, ordinary formulas for estimating the standard error of sample statistics will produce values that are too small.

Before attempting any analysis of NAEP data, users should understand the special characteristics of the NAEP sampling design (Chapters 2 and 4). Alternate methods for computing standard errors and recommended formulas for obtaining degrees of freedom are given in Chapter 8.

Chapter 2

SPECIAL CONSIDERATIONS FOR USERS

Chapter 2: SPECIAL CONSIDERATIONS FOR USERS

Because of the complexity of the NAEP design, it is important for users to have some understanding of it before performing analyses of the data. The following sections highlight areas of potential importance to the user in constructing analyses.

Details of the design and data analysis for the 1990 Trial State Assessment are provided in *The Technical Report of NAEP's 1990 Trial State Assessment Program* (Koffler, 1991).

2.1 THE NATIONAL COMPARISON SAMPLE OF STUDENTS

One of the purposes of the Trial State Assessment Program was to allow each participating state to compare its results with those of the nation as a whole and with those of the geographic region in which that state is located.[1] To permit such comparisons, a nationally representative sample of eighth-grade students was assessed as part of the national assessment using the same instruments used in the Trial State Assessment.

Because of differences between the state and national samples (described in Chapter 8), it was necessary to create a subsample from the full national sample to allow for valid state/nation comparisons. Data from this subsample (referred to as the national winter public-school sample, or "NWP" sample) are included on the secondary-use data files, along with the appropriate weights to be used for analyses. Chapter 8 provides information on conducting analyses using the NWP sample.

2.2 FOCUSED-BIB SPIRAL METHOD OF ADMINISTRATION

The term "focused-BIB spiral" refers to the method used to assemble assessment items into instruments. This method was developed to allow the study of the interrelationships among all items within a subject area. As a result of this design, all items are given to approximately the same number of students, but no student receives all items.

The focused-BIB spiral design for the mathematics booklets in the Trial State Assessment is discussed in Chapter 3.

2.3 REPORTING SUBGROUPS AND OTHER VARIABLES

In addition to reporting overall state or national achievement results, NAEP reports results for several student subgroups -- gender, race/ethnicity, type of community, and level of parents' education. Some of these subgroups were derived from students' responses to one or more assessment items. Chapter 6 defines and explains the reporting subgroups.

---

[1] No regions have been designated for the territories.

Certain derived variables on the data files were created through the systematic combination of values from one or more items from the student, teacher, or school questionnaire. The derived variables are described in Chapter 6.

The files also contain mathematics proficiency variables, called plausible values. These variables, developed for scaling purposes, are described in Chapter 6; their explanation and use are given in Chapter 7.

Some variables on the files were taken from sources other than the assessment instruments. For optimal use of these variables, see their explanations in Chapter 6.

2.4 RESPONSE DATA FROM TEACHERS

The mathematics teachers of the students assessed in both the national mathematics assessment and the Trial State Assessment were asked to complete a two-part questionnaire about their instructional practices, teaching backgrounds, and other characteristics. The first part of the questionnaire pertained to the teachers' background and training; the second pertained to the programs and instructional methods the teacher used for each class containing an assessed student.

In the NAEP data files, the data from the teacher questionnaire have already been linked with the appropriate student response data and included on the student data records, allowing correct and efficient analysis of the teacher/student data without requiring users to match data from separate files.

Note: The purpose of this sample is to estimate the numbers of *students* whose teachers have various attributes, not to estimate the attributes of the teacher population. Because of the nature of the sampling for the Trial State Assessment, the responses to the mathematics teacher questionnaire do not necessarily represent all eighth-grade mathematics teachers in a state. Rather, they represent the teachers of the particular students being assessed.

2.5 USING WEIGHTS

In the NAEP sampling design, students do not have an equal probability of being selected. Therefore, as in all such complex surveys, each student has been assigned a sampling weight. When computing descriptive statistics or conducting inferential procedures, one should weight the data properly for each student. Performing statistical analyses without weights can lead to misleading results.

Chapter 4 explains the weight variables and how they were developed; Chapter 8 explains how to use weights in performing analyses.

2.6 ERROR ESTIMATION

The 1990 NAEP sampling design involved the selection of clusters of students from the same school, as well as clusters of schools from urbanicity, income, and minority strata (in the case of the Trial State Assessment) and from the same geographically defined primary sampling unit, or PSU (in the case of the national assessment). As a result, observations are not independent of one another as they are in a simple random sample. Therefore, use of ordinary formulas for estimating the standard error of sample statistics will result in values that are too small. Alternate methods of computing standard errors are provided in Chapter 8.

Another effect of the sampling design is a reduction of the effective degrees of freedom, which in the 1990 NAEP design are a function of the number of clusters of schools (for the Trial State Assessment) or clusters of PSUs (for the national assessment) and the number of strata in the design, rather than the number of subjects. Recommended formulas for obtaining degrees of freedom can be found in Chapter 8.

2.7 MONITORED AND UNMONITORED ASSESSMENT SESSIONS

As part of the effort to ensure security and uniformity in the administration of the Trial State Assessment, a random half of the assessment sessions were monitored by trained quality control monitors. Within each state, and across all states, randomly equivalent samples of students received each block of cognitive items in a particular position within a booklet under monitored and unmonitored administration conditions. Thus, it was possible to conduct analyses comparing the data from the monitored sessions with the data from the unmonitored sessions.

Special weights are provided on the data files for comparing samples of students in the monitored and unmonitored sessions. Chapter 8 describes the use of these weights for analyses.

Chapter 3

INSTRUMENT DESIGN

Chapter 3: INSTRUMENT DESIGN

3.1 INTRODUCTION

In the 1990 Trial State Assessment, several types of instruments were used to collect data about students, teachers, and schools. Each assessed student received a booklet containing three segments of cognitive mathematics items, a demographic questionnaire, and a mathematics background questionnaire. An excluded student questionnaire was completed by school officials for each sampled student who was deemed unable to take part in the assessment. Teacher questionnaires were given to the mathematics teachers of the assessed students. A school characteristics and policies questionnaire was distributed to each participating school.

This chapter describes the content and organization of the assessment instruments. See Chapter 4 for information about how schools, students, and teachers were selected to participate in the assessment.

3.2 STUDENT ASSESSMENT BOOKLETS

3.2.1 Booklet Content

The mathematics items for the 1990 assessment were developed from a set of objectives created through a broad-based consensus process. The framework adopted for the 1990 mathematics assessment was organized according to three mathematical abilities and five content areas. The mathematical abilities assessed were conceptual understanding, procedural knowledge, and problem solving. Content was drawn primarily from elementary and secondary school mathematics up to, but not including, calculus. The content areas assessed were numbers and operations; measurement; geometry; data analysis, statistics, and probability; and algebra and functions.

The overall pool of items for the Trial State Assessment consisted of 137 items -- 102 multiple-choice and 35 open-ended items -- that were designed to provide an extended view of students' mathematical knowledge and skills. Table 3-1 provides the number of items for each content and ability group included in the Trial State Assessment. These same items were also used in the national mathematics assessment.

The 137 cognitive mathematics items were assembled into seven different 15-minute segments or "blocks." Two of the seven blocks were designed to be answered using a calculator and one using a protractor/ruler. The blocks were assembled three to a booklet, and each student was asked to respond to one booklet.

Table 3-1
Content-by-Ability Distribution of Items
Grade 8

Mathematical Abilities	Content Areas					
	Numbers and Operations	Measurement	Geometry	Data Analysis, Statistics, and Probability	Algebra and Functions	TOTAL
Conceptual Understanding	18	7	13	9	12	59
Procedural Knowledge	15	9	4	5	8	41
Problem Solving	12	5	9	5	6	37
TOTAL	45	21	26	19	26	137

In addition to the cognitive item blocks, each booklet contained a set of questions about the students' demographic characteristics and a set of questions about the students' mathematics background. These questionnaires are described in section 3.3.1.

3.2.2 Booklet Assembly

The assembly of mathematics items into booklets and their subsequent assignment to assessed students was determined by a *balanced incomplete block* (BIB) design with *spiraled* administration.

The first step in implementing BIB spiraling required dividing the total pool of mathematics items into blocks (labeled M3 through M9) designed to take 15 minutes to complete. These blocks were then assembled into booklets containing a 5-minute demographics questionnaire block (B1), a 5-minute mathematics background block (M2), and three blocks of mathematics items. Thus, the overall assessment time for each student was approximately 55 minutes. The mathematics blocks were assigned to booklets in such a way that each cognitive item block appeared in the same number of booklets and every pair of blocks appeared together in exactly one booklet. This is the *balanced* part of the balanced incomplete block design. It is an *incomplete* block design because no booklet contained all items and hence there is *incomplete* data for each assessed student.

The BIB design for the 1990 national mathematics assessment (and, therefore, for the Trial State Assessment) was *focused* -- each block was paired with every other mathematics block but not with blocks from other subject areas. The *focused*-BIB design also balances the order of presentation of the blocks of items -- every block appears as the first cognitive block in one booklet, as the second block in another booklet, and as the third block in a third booklet.

The focused-BIB design used in 1990 required that seven blocks of mathematics items be assembled into seven booklets. The assessment booklets were then *spiraled* and bundled.

Spiraling involves interleaving the booklets in a systematic sequence so that each booklet appears an appropriate number of times in the sample. The bundles were designed so that each booklet would appear equally often in each position in a bundle.

The final step in the BIB-spiraling procedure was the assigning of the booklets to the assessed students. The students within an assessment session were assigned booklets in the order in which the booklets were bundled. Thus, students in an assessment session received different booklets, and only several students in the session received the same booklet. In the Trial State Assessment BIB-spiral design, representative and randomly equivalent samples of about 2,500 students responded to each item.

Table 3-2 shows the composition of each block of items administered in the Trial State Assessment Program. Table 3-3 shows which blocks were contained in each booklet and how the focused-BIB design was used to combine the seven cognitive blocks into seven booklets. Note that these same blocks and focused-BIB design also were used for the eighth-grade national assessment.

3.2.3 Release Status for Item Blocks

As described in Chapter 1, some NAEP cognitive items are classified as *non-released.* These are items that are reserved exclusively for NAEP use -- for example, for administration in future assessments to permit analysis of trends in performance levels. In the 1990 Trial State Assessment data files, four blocks of cognitive mathematics items -- blocks M4, M5, M6, and M8 -- are classified as non-released. In the data files and codebooks, the item stem for these exercises has been replaced with a short description and the distractors have been withheld.

Items in the remaining three cognitive blocks -- M3, M7, and M9 -- and all of the items in the noncognitive blocks and the questionnaires are classified as public-release.

3.3 QUESTIONNAIRES

As part of the Trial State Assessment (as well as the national assessment), a series of questionnaires was used to collect information about assessed students, excluded students, mathematics teachers, and schools. The questionnaires are described in the following sections; sampling methods are described in Chapter 4.

3.3.1 Student Questionnaires

In addition to the cognitive questions, the booklets used in the 1990 Trial State Assessment included two 5-minute sets of general demographic and mathematics background questions designed to gather contextual information about students, their experiences in mathematics, and their attitudes toward the subject.

Table 3-2

Cognitive and Noncognitive Block Information

Block	Type	Total Number of Items	Number of Multiple-Choice Items	Number of Open-ended Items	Booklets Containing Block
B1	Common Background	22	22	0	8 - 14
M2	Mathematics Background	22	22	0	8 - 14
M3	Mathematics Cognitive	23	19	4	8, 12, 14
M4	Mathematics Cognitive	21	21	0	8, 9, 13
M5	Mathematics Cognitive	16	0	16	9, 10, 14
M6	Mathematics Cognitive (Protractor/Ruler)	21	16	5	8, 10, 11
M7	Mathematics Cognitive	18	17	1	9, 11, 12
M8	Mathematics Cognitive (Calculator)	18	16	2	10, 12, 13
M9	Mathematics Cognitive (Calculator)	20	13	7	11, 13, 14

Table 3-3

Contents of Assessment Booklets

Booklet Number	Common Background Block	Mathematics Background Block	Cognitive Blocks		
8	B1	M2	M3	M4	M6[1]
9	B1	M2	M4	M5	M7
10	B1	M2	M5	M6[1]	M8[2]
11	B1	M2	M6[1]	M7	M9[2]
12	B1	M2	M7	M8[2]	M3
13	B1	M2	M8[2]	M9[2]	M4
14	B1	M2	M9[2]	M3	M5

---

[1] Protractor/ruler needed for this block

[2] Calculator needed for this block

The **student demographics questionnaire** (block B1 -- 22 questions) included questions about race/ethnicity, language spoken in the home, mother's and father's level of education, reading materials in the home, homework, attendance, school climate, academic expectations, which parents live at home, and which parents work. This questionnaire was the first section in every booklet. In many cases the questions used were continued from prior assessments.

Three categories of information were represented in the second 5-minute section of background questions called the **student mathematics questionnaire** (block M2 -- 22 questions):

- Time Spent Studying Mathematics: Students were asked to describe both the amount of instruction they received in mathematics and the time spent on mathematics homework.
- Instructional Practices: Students were asked to report their experience in using various instructional materials in the mathematics classroom, including calculators, models, and manipulatives. In addition, they were asked about the instructional practices of their mathematics teachers and the extent to which the students themselves practiced the communication of mathematical ideas--such as writing out explanations, justifications, or proofs--in their mathematics classes.
- Attitudes Towards Mathematics: Students were asked a series of questions about their attitudes and perceptions about mathematics, such as whether they enjoyed mathematics and whether they were good in mathematics.

Data from these questionnaires are contained on the student data files.

3.3.2 Excluded Student Questionnaire

The **Excluded Student Questionnaire** was completed by the teachers of those students who were selected to participate in the Trial State Assessment but were determined by the school to be ineligible to be assessed because they either had an Individualized Education Plan (IEP) and were not mainstreamed at least 50 percent of the time, or were categorized as Limited English Proficient (LEP).

The questionnaire contained 27 questions about the characteristics of the student and the reason for exclusion. For students with an Individual Education Plan, the questionnaire included questions about the student's functional grade level, mainstreaming, and special education programs. For Limited English Proficient students, it asked about the student's native language, time spent in special education and language programs, and the level of the student's English language proficiency.

Information from this questionnaire is contained in the excluded student data files.

3.3.3 The Teacher Questionnaire

To supplement the information on instruction reported by students, the mathematics teachers of the eighth graders participating in the Trial State Assessment were asked to complete a questionnaire about their instructional practices, teaching backgrounds, and characteristics. The teacher questionnaire contained two parts. The first part pertained to the teachers' background and training. The second part pertained to the procedures the teacher uses for *each class* containing an assessed student.

The Teacher Questionnaire, Part I: Background and Training (34 questions) included questions pertaining to gender, race/ethnicity, years of teaching experience, certification, degrees, major and minor fields study, coursework in education, coursework in subject area, in-service training, extent of control over classroom, instruction, and curriculum, and availability of resources for classroom.

The **Teacher Questionnaire, Part II: Classroom by Classroom Information** (35 questions) included questions on the ability level of students in the class, whether students were assigned to the class by ability level, time on task, homework assignments, frequency of instructional activities used in class, instructional emphasis given to the topics and skills covered in the assessment, and use of particular resources.

Data collected from the teacher questionnaires are appended to the appropriate student records in the student data files.

Note: The purpose of this sample is to estimate the numbers of *students* whose teachers have various attributes, not to estimate the attributes of the teacher population. Because of the nature of the sampling for the Trial State Assessment, the responses to the mathematics teacher questionnaire do not necessarily represent all eighth-grade mathematics teachers in a state. Rather, they represent the teachers of the particular students being assessed.

3.3.4 School Questionnaire

A **School Characteristics and Policies Questionnaire** was given to the principal or other administrator of each school that participated in the Trial State Assessment Program. This questionnaire included 117 questions about background and characteristics of school principals, length of school day and year, school enrollment, absenteeism, drop-out rates, size and composition of teaching staff, policies about tracking, curriculum, testing practices and use, special priorities and school-wide programs, availability of resources, special services, community services, policies for parental involvement, and school-wide problems.

Data collected from the school questionnaire can be found in the school data files.

Chapter 4

SAMPLE SELECTION AND WEIGHTS

Chapter 4: SAMPLE SELECTION AND WEIGHTS

4.1 INTRODUCTION

This chapter describes the methods used by Westat, Inc., the survey subcontractor, to select the samples for the states participating in the 1990 Trial State Assessment (section 4.2) and provides an overview of the sampling weights on the data files and how they were derived for the state samples (section 4.3). A discussion of how to use the sampling weights is given in Chapter 8. Sampling and weighting procedures for the national portion of the assessment are described in the technical report for the 1990 national assessment.

4.2 SAMPLE SELECTION

The representative sample of eighth-grade students assessed in the Trial State Assessment came from about 100 public schools in each jurisdiction, unless a jurisdiction had fewer than 100 public schools with eighth-grade students, in which case all or almost all such schools participated. The sample of schools in each state was selected with probability proportionate to size, where the measure of size was equal to the number of students enrolled in the eighth grade in each school. The school samples were implicitly stratified based on urbanicity, percentage of minority enrollment, and household income.

Except for some schools in a few states, schools selected for the 1990 national assessment for age 13/grade 8 were excluded from the Trial State Assessment. Appropriate weighting adjustments were used to ensure that these exclusions did not introduce bias into estimates from the state samples. The goal was 100 percent participation of all selected schools. Many of the schools that declined to participate were replaced in the sample by substitute selections.

The target population for the Trial State Assessment Program consisted of eighth-grade students enrolled in public schools. In general, slightly more than 100 schools per state were selected to allow for the fact that some selected schools would not have any eligible students enrolled. Such schools arose as a result of errors in the list of schools used to compile the sampling frame. Thirty students selected from each school provided a sample size of approximately 3,000 students per state. The student sample size of 30 for each school was chosen to ensure at least 2,000 students participating from each state, accounting for school nonresponse, exclusion of students, inaccuracies in the measures of enrollment, and student absenteeism from the assessment.

The levels of school participation varied considerably across the 40 participating jurisdictions. Prior to substitution, weighted response rates (for which each school was weighted in proportion to the size of the student population represented by that school in the sample) ranged from 73 percent to 100 percent. The two states with relative low initial response rates obtained good cooperation from their substitute schools, so that, after substitution, the lowest response rate was 85 percent.

Student participation rates were uniformly high, except in one state where parental consent requirements kept this rate at 80 percent.

Details of school and student participation rates and the rate of student exclusion for each of the 40 participating jurisdictions are given in *The Technical Report of NAEP's 1990 Trial State Assessment.*

The schools within each state were stratified by the following variables:

- Urbanicity (central city, suburban, other)
- Percentage of Black and Hispanic students enrolled
- Median household income

All states, except for those with 100 schools or fewer, were stratified by urbanicity and income variables. Only states with significant minority populations were stratified based on minority enrollment.

In contrast to the national assessment, which was administered by Westat field personnel, the Trial State Assessment was administered by local school or district personnel. To check on the consistency of assessment administration conditions, half the schools in the sample were monitored by Westat field staff and half were unmonitored, to permit comparisons between the two. The sample in each state was designed both to produce aggregate estimates for the state, and various subpopulations (depending upon the size and distribution of the various subpopulations within the state), and also to enable comparisons to be made, at the state level, between administration with monitoring and without monitoring.

The following sections provide some details of the various aspects of selecting the sample for the 1990 Trial State Assessment, including frame construction, the stratification process, updating the school frame with new schools, and the actual sample selection. A fuller discussion of sample selection is given in *The Technical Report of NAEP's 1990 Trial State Assessment.*

4.2.1 Selection of Schools

4.2.1.1 Frame Construction

Three sources of data were combined to construct the school sampling frame:

- The NCES Common Core of Data for 1988.

- Data on school-level minority enrollment collected from school districts during the sample design phase of the 1988 and 1990 NAEP samples.

- 1980 Census data broken down to the ZIP code level, as provided by Donnelley Marketing Information Services.

For the school-level sample design, the frame variables used were total enrollment, eighth-grade enrollment, urbanicity, minority enrollment, and median household income.

In order to minimize overlap with the national NAEP school samples, in general schools selected for the national sample were excluded from the state frame. Weighting adjustments were made to account for this procedure and render unbiased estimates. A similar approach was used to exclude those schools having both grades 8 and 10 that were included in the school sample for the 1990 National Educational Longitudinal Study First Phase Follow-Up.

4.2.1.2 Stratification

States were stratified on urbanicity, percentage of minority enrollment, and household income depending on the number of eighth-grade schools within the state and the percentage of minority students within each urbanicity class:

- In states with 105 schools or less, schools were not stratified at all, since all schools in these states with at least 20 students were selected for the assessment. If a sample of smaller schools was drawn, rather than selecting them all, then this sample was not stratified. Schools in these states were called Type 1 clusters.

- In states which *either* had 106 to 200 schools *or* a low percentage of minority students, schools were stratified by urbanicity and household income.

- In states which had more than 200 schools *and* a high percentage of minority students, schools were stratified by urbanicity, percentage of minority enrollment, and household income.

- In those states with high percentages of both Black and Hispanic students and more than 200 schools, schools were stratified on the basis of percentage of Black enrollment and percentage of Hispanic enrollment.

Urbanicity was categorized as *Central City*, *Suburban*, and *Other*, although these classes were collapsed in some cases. If any urbanicity class had more than 10 percent Black students or 7 percent Hispanic students but not more than 20 percent of both, the schools within the urbanicity class were stratified by ordering the schools by the percentage of minority enrollment and dividing the schools into three groups with an approximately equal number of schools in each. Urbanicity classes with fewer than 10 percent Black students and 7 percent Hispanic students were not stratified by minority enrollment. Where there were high percentages of both Black and Hispanic students (i.e., more than 20 percent of each), four strata were formed:

- High Black/high Hispanic: schools above the medians for both percentage of Black students and percentage of Hispanic students.

- High Black/low Hispanic: schools above the median for percentage of Black students but below the median for percentage of Hispanic students.

- **Low Black/high Hispanic:** schools below the median for percentage of Black students but above the median for percentage of Hispanic students.

- **Low Black/low Hispanic:** schools below the medians for both percentage of Black students and percentage of Hispanic students.

Within these classes defined by urbanicity and minority enrollment, schools were sorted in serpentine order by the median household income so that bordering schools in different classes would be the most similar. For instance, within the suburban urbanicity, if the low minority class was sorted from highest median income to lowest, then the intermediate minority class was sorted from lowest median income to highest, and the highest minority class was sorted from highest median income to lowest.

Schools with 19 or Fewer Students

Since the target assessment size for each school was about 25 after allowance for exclusion and absenteeism, schools with 19 or fewer eighth-grade students were handled by one of two different methods, depending on the prevalence of these schools within the given state, and of the students attending them. These special procedures were adopted to provide control over the sample sizes of both schools and students and are referred to as "geographic" and "stratified" grouping.

Geographic Grouping: In states with a relatively small number of such schools (specifically, fewer than 20 percent of the schools for the state, with fewer than 1 percent of the total eighth-grade students), small schools were grouped geographically with larger ones (eighth-grade enrollment of 20 or more), and then the resulting pairs (or possibly larger groups) were initially sampled together as a single unit. These units were called Type 2 clusters. Data for stratification were pooled between the paired schools.

Stratified Grouping: In states with larger numbers of small schools, schools were stratified into two groups, depending on whether or not their eighth-grade enrollment was 20 or more. Schools whose eighth-grade enrollment was at least 20 were referred to as Type 3A clusters. Schools with fewer than 20 eighth-grade students were clustered into groups called Type 3B clusters. This approach assured that no clusters had student enrollment less than 20.

The number of Type 3B clusters selected was proportionate to the number of students that attended schools with 20 or fewer students, up to a maximum of ten clusters. This maximum was imposed to keep the size of the sample of small schools to within reasonable bounds. These Type 3B clusters were not stratified on urbanicity, minority enrollment, or income, but were selected systematically with probability proportionate to the total eighth-grade enrollment in the cluster. Type 3A clusters were stratified on urbanicity, minority enrollment, and income as discussed above.

4.2.1.3 Selection of School Sample

States with Geographic Clustering of Small Schools (Type 2 Clusters)

In states with 200 or fewer schools, clusters were sorted by urbanicity and median income. In states with more than 200 schools, clusters were sorted by urbanicity, minority strata (which varied by state and urbanicity level), and median income. After the removal of certainty schools (those with selection probability greater than 1), a systematic sample of clusters was then selected with probability proportionate to total eighth-grade enrollment, to provide a total sample of 105 clusters.

Following the selection of clusters, there was some thinning of small schools. The purpose of thinning was to give students in small schools (enrollment less than 20) approximately the same chance of selection as those from larger schools, and to control the sample size of schools to be close to the desired number of 105. All small schools in a cluster were discarded from the sample with probability 30/X, where X denotes the total enrollment for all schools in the cluster to which the small schools belonged. Otherwise, the small schools were retained in the sample.

States with Stratification of Small Schools (Type 3A and 3B Clusters)

For all states, the percentage of eighth-grade students in the state who attended small schools (i.e., schools with 19 or fewer students) was determined. The sample design for the selection of *small schools* was the same for all such states, except Montana, North Dakota, and Nebraska. In every state the percent of students in small schools, *p*, was rounded to the nearest integer that was at least one, with this integer being called *k*. Montana, North Dakota, and Nebraska were exceptions where the values of *k* which were in excess of 10 in each case were reduced to 10 to keep the sample size of small schools to within reasonable bounds. A random sample of *k* clusters of small schools was selected.

The sample selection of *large schools* (i.e., schools with 20 or more students) varied by state. In states with 105 or fewer schools, all large schools were selected. In states with 106 to 200 schools, after the large schools were sorted by urbanicity and median income, and certainty selections were removed, a systematic sample of schools was selected with probability proportionate to total eighth-grade enrollment, such that the total sample size of large schools, including certainty selections, was (105 - *k*). The exceptions were Montana, North Dakota, and Nebraska, where the total sample size for large schools was set at 90 in each case. Once again, the special exception for these states was designed to limit the total number of schools selected. In states with more than 200 schools, the large schools were sorted by urbanicity, minority strata, and median income, and the certainty schools were removed. Then a systematic sample of schools was selected with probability proportionate to total eighth-grade enrollment such that the total sample of large schools, including certainties, was (105 - *k*).

After the sample of schools was selected, weighted tabulations were produced to verify that it was representative of the population. The number of clusters sampled, along with the estimated and actual counts of clusters and students, were listed by urbanicity level and minority level for each state. The differences between the actual and estimated numbers of clusters were also calculated.

Designating Schools to be Monitored

The objective in assigning each school to be monitored or unmonitored was to produce two equivalent half-samples. This was achieved by pairing similar clusters, and randomly designating one pair member to be monitored, independently from pair to pair. For Type 3B clusters, the procedure was applied to schools within clusters, with random sort order within cluster.

This procedure was followed for all schools in all states except for those states where schools were not stratified. For these states, pairing was done on the basis of school size, because these states in general showed little variation with regard to urbanicity and race/ethnicity and no household income data were readily available.

Updating the School Sample with New Schools

In sampling for the Trial State Assessment, some districts had new schools that were not listed on the sampling frame, either because these schools were completely new or because they had been formed by some combination of old schools. In either case, to provide a mechanism for allowing these new schools into the sample after the initial sample was selected, all districts in which schools were sampled were contacted and provided with the list of schools from the sampling frame for that district. The district was then asked to provide an updated list containing any schools not listed on the frame which were operating in the 1989-1990 school year and which contained the eligible grade. A sample of new schools was then selected from the lists provided. In order that all schools in each participating state had a chance of being selected for the Trial State Assessment, schools on the updated list were sampled and, if selected, were asked to participate in the program. Since a self-weighting sample of students was desired, the required sample size of new schools depended on the method used to weight the data estimation. The determination as to how many new schools were selected and how the data from selected schools were weighted is discussed in *The Technical Report of NAEP's 1990 Trial State Assessment*.

School Substitution

A substitute school was selected for each selected school containing eligible students, for which school nonparticipation was established by the state coordinator as of November 10, 1989. The process of selecting a substitute for a school involved identifying the most similar school in terms of the following characteristics: urbanicity, percentage of Black enrollment, percentage of Hispanic enrollment, eighth-grade enrollment, and median income.

To identify candidates for substitution, a set of schools were found which provided reasonable matches with regard to eighth-grade enrollment and percentage of Black and Hispanic enrollment. From among this set a match was selected considering all five characteristics. Schools in the national assessment sample or the 1990 National Education Longitudinal Study First Phase Follow-up were avoided in the selection of substitutes, where possible. Furthermore, the substitute was selected from the same district, wherever possible, to avoid placing the burden of replacing a refusing school from one district on another district. This was often not possible, however, because in the majority of cases, the decision not to participate was made at the district level.

In a few cases where no suitable substitute could be found among those schools not sampled (most often because all or most schools were included in the original sample), a school already in the sample conducted a double session, of which one session served as a substitute for students in the refusing school. The same criteria were applied in selecting the schools that conducted double sessions, i.e., a reasonable match was found based on eighth-grade enrollment, percentage of Black and Hispanic enrollment, median income, and urbanicity.

4.2.2 Selection of Student Samples

For all schools in each state, a student sample size of 30 was drawn from each selected school per state, expect for states with fewer than 100 schools (Type 1 clusters). In these states either 60 or 100 students were sampled in the larger-certainty schools, depending on the size of the state and the size of the school.

In November 1989, school officials were asked to forward a list of the names of all of the eighth-grade students enrolled in the school to a central location (usually the State Department of Education). Schools were not asked to list students in any particular order, but were asked to implement checks to ensure that all eighth-grade students were listed. Based on the total number of students on this list, called the Student Listing Form, sample line numbers were generated for student sample selection. To generate these line numbers, the person responsible for drawing the sample (typically, the State Supervisor) went to the State Department of Education and entered the following into a calculator that had been programmed with the sampling algorithm: the number of students on the Student Listing Form, the state identity, and the sample size if it was different from 30. The calculator generated a random start which was used to systematically select the 30 (or more if necessary) line numbers. To compensate for new enrollees not on the form, extra line numbers were generated for a supplemental student sample of new students. All students were selected in those schools with 35 or fewer eighth-grade enrollees. This sample design was intended to give each student within the state approximately the same probability of selection.

After the student sample was selected, the administrator at each school excluded students who were incapable of taking the assessment -- a subset of those students who had an Individualized Education Plan or who were Limited English Proficient.

When the assessment was conducted in a given school, a count was made of the number of non-excluded students who did not attend the session. If this number exceeded three students, the school was instructed to conduct a make-up session, to which all students who were absent from the initial session were asked to attend.

4.3 WEIGHTING PROCEDURES

Following the collection of assessment and background data from and about assessed and excluded students, the processes of deriving sampling weights and associated sets of replicate weights were carried out. The sampling weights are needed to make valid inferences from the student samples to the respective populations from which they were drawn and should be used for all analyses, whether exploratory or confirmatory. Replicate weights were used in the estimation of sampling variance, through the procedure known as jackknife repeated replication. See Chapter 8 for information about how to use the sampling and replicate weights.

The following is an overview describing the weight variables on the data files and summarizing the methods use to derive them. Full details of the weighting procedures are given in *The Technical Report of NAEP's 1990 Trial State Assessment.*

4.3.1 Full-sample Weights

Each student was assigned a weight, to be used for making inferences about the state's students, without regard to whether the student was in a monitored or unmonitored session. This weight is known as the full, or overall, sample weight.

The student full-sample weight contained three components -- a base weight, an adjustment for school nonparticipation, and an adjustment for student nonparticipation. These are described in a general way below; full details are given in *The Technical Report of NAEP's 1990 Trial State Assessment.*

The student base weight -- the inverse of the overall probability of selection of the sampled student -- incorporated the probability of selection of the student's school, and of the student within school, and accounts for the impact of procedures used to keep to a minimum the overlap of the school sample with both the national assessment eighth-grade school sample, and the sample of schools involved in the National Educational Longitudinal Study First Phase Follow-up. The student base weight was a product of the base weight of the school in which the student was enrolled and the within-school student weight (STUDWGT), where the student weight was given as:

$$\textit{Student Weight} = \frac{\textit{Actual Eighth-grade Enrollment}}{\textit{Sample Size}} ,$$

reflecting the within-school student probability of selection.

The base weight was then adjusted for two sources of nonparticipation -- school level and student level. These weighting adjustments seek to reduce the potential for bias from such nonparticipation by increasing the weights of students from schools similar to those schools not participating (adjustment factor ADJFAC), and increasing the weights of students similar to those students from within participating schools who did not attend the assessment session (or a make-up) as scheduled (adjustment factor STUDNRF).

For excluded students, the base weight and school nonparticipation adjustment factor were the same as for assessed students from the same school. Excluded student nonresponse adjustments were calculated to account for the fact that an excluded student questionnaire was not returned for a small percentage of excluded students.

Either of two alternatively scaled weights can be used as the full-sample weight for analyses at the student level. The first of these, ORIGWT, has been scaled so that the sum of weights for all students in each state estimates the total number of eighth-grade assessable students in that state's public schools. The second of these, WEIGHT, is a proportional rescaling of ORIGWT, carried out so that the sum of WEIGHT across students and states is equal to the total Trial State Assessment sample size across all states (i.e., the total number of assessed students in the Trial State Assessment). Both weights should provide identical estimates of means, proportions, correlations, and other statistics of interest in analyses within each state as well as analyses involving data from more than one state.

The base weight assigned to a school (BASEWT) was the reciprocal of the probability of selection of that school. The base weight reflected the actual probability used to select the school from the frame. It also included two factors that reflected the impact on the selection probability of the avoidance of school sample overlap with both the national NAEP samples and the 1990 National Educational Longitudinal Study first follow-up of tenth-grade students. Schools that substituted for a refusing school were assigned the weight of the refusing school, unless of course the substitute refused. For schools that conducted double sessions because they were substitutes for a refusing school, half of the students were assigned the school base weight of the participating school and half were assigned the weight of the refusing school. These half samples were chosen at random, with each half-sample constituting a simple random sub-sample of the full sample of students from the schools.

The final school weight, adjusted for nonparticipation, is named SCHWTF. This weight should be used in analyses of the school questionnaire data.

4.3.2 Comparison Weights for Monitored/Nonmonitored Sessions

A second student weight, known as the comparison weight, was derived from the full sample weight for use in making comparisons, within and across states, in the performance of students who were assessed in monitored sessions with those assessed in unmonitored sessions. The comparison weights were obtained from the full-sample weights using an additional adjustment procedure, described as follows:

The monitored and the unmonitored schools comprised two random half-samples. In order to compare the results for them with greater precision, a form of poststratification to certain characteristics of the whole sample was employed on each of these half samples. This procedure, called raking (adjustment factor RAKADJ), was intended to reduce variance of estimates of differences between monitored and unmonitored sessions, by controlling for sampling variability in student characteristics known to be related to mathematics proficiency, but unrelated to the monitoring process. This adjustment (RAKADJ) resulted in the distributions of weighted counts for various student characteristics that were very similar for the students from the monitored sessions and the unmonitored sessions.

As with the full-sample weight, two alternative scalings are available, CWEIGHT (which sums to the overall sample size) and CORIGWT (which sums to population sizes). Either CWEIGHT or CORIGWT should be used in lieu of WEIGHT or ORIGWT for all analyses intended to compare statistics (such as a mean, proportion, or correlation) obtained from monitored sessions to the same statistic obtained in the unmonitored sessions.

4.3.3 Replicate Weights

In addition to estimation weights, a set of replicate weights was provided for each student. These replicate weights are used in estimating sampling errors of estimates obtained from the data, using the jackknife repeated replication (or jackknifing) method. Chapter 8 describes the method of using these replicate weights to estimate sampling errors. The methods of deriving these weights were aimed at reflecting appropriately the features of the sample design in each state so that when the jackknife variance estimation procedure was implemented as intended, approximately unbiased estimates of sampling variance would result.

Replication estimates the variance of the full sample. This process involves repeatedly selecting portions of the sample to calculate the estimate of interest. The estimates that result are called replicate estimates. The variability among these calculated quantities is used to obtain the full sample variance. The process of forming these replicate estimates involves first dividing the sample elements among a set of replicate groups, then using the pattern of replicate groups in a systematic fashion to apply replicate weights to the file.

Similar to the estimation weights, two sets of replicate weights were derived for each student. The first set is referred to as the overall replicate weights (SRWT01-56), and correspond to the full sample weight (WEIGHT). They are used for estimating the sampling errors of estimates derived using the full sample weights. These weights are designed to reflect the method of sampling schools, and account for the type of stratification used and whether or not the student's school was included in the sample with certainty. The method of sampling students within schools is also reflected, implicitly in the case of noncertainty schools and explicitly for schools included with certainty. These overall replicate weights also reflect the impact on sampling errors of the school- and student-level nonresponse adjustments applied to the full sample weights.

The second set of replicate weights, known as the comparison replicate weights (CSRWT01-56), are for use in estimating sampling errors of estimates obtained using the comparison weights (CWEIGHT). These replicate weights differ from the overall replicate weights in two ways. First, in addition to reflecting features of the sample design and weighting procedures, they reflect the impact on sampling error of the raking procedure (see section 4.3.2) used to equate weighted distributions from the monitored and unmonitored half samples in each state. Second, in those states where some or all schools were selected into the sample with certainty, the comparison weights reflect the fact that such certainty selections were assigned to be monitored or unmonitored at random. Thus, these certainty schools contribute a school level component of variance to the comparison of monitored and unmonitored assessments, which is appropriately reflected in the comparison replicate weights.

At the school level, the replicate weights SCHWT01-56 on the school data files should be used to estimate the variance for population estimates obtained using the school weight (SCHWTF).

4.3.4 Summary of Weights and Their Use

Table 4-1 gives a summary of the sample weights and replicate weights and the purposes for which they should be used. Chapter 8 provides a detailed discussion of how to use the weights in conducting analyses.

Table 4-1

Summary of Weights for the 1990 Trial State Assessment

Group	Sample Weight	Replicate Weights	Use
Assessed Students (1)	WEIGHT	SRWT01-56	Student-level analyses comparing students within or across states Student-level analyses comparing students in state to students in nation
Assessed Students (2)	CWEIGHT	CSRWT01-56	Student-level analyses comparing students within or across states when comparing monitored and unmonitored sessions
Excluded Students	XWEIGHT	XRWT01-56	Excluded student analyses within or across states
Schools	SCHWTF	SCHWT01-56	School-level analyses within or across states School-level analyses between nation and states

Chapter 5

DATA COLLECTION, MATERIALS PROCESSING, PROFESSIONAL SCORING, AND DATABASE CREATION

Chapter 5: DATA COLLECTION, MATERIALS PROCESSING, PROFESSIONAL SCORING, AND DATABASE CREATION

5.1 INTRODUCTION

In addition to sample selection, Westat, Inc., was responsible for field administration and data collection for the 1990 Trial State Assessment. When data collection was completed, assessment instruments were sent to National Computer Systems for processing and scoring. The resulting data files were then sent to ETS, where they were transcribed to a database ready for analysis. This chapter provides an overview of these activities, which are described in detail in *The Technical Report of NAEP's 1990 Trial State Assessment Program.*

5.2 DATA COLLECTION AND FIELD ADMINISTRATION

Data collection for the 1990 Trial State Assessment involved a collaborative effort between staff in the participating states and schools and the NAEP contractors, especially Westat, Inc., the field administration contractor. Between February 5 and March 2, 1990, Westat sampled a total of over 100,000 students from more than 3,500 schools across the 40 participating states. Westat's data collection responsibilities included developing administration procedures and manuals, training the state personnel who conducted the assessments, and conducting an extensive quality assurance program.

Each state participating in the 1990 Trial State Assessment was asked to appoint a state coordinator who became the liaison between NAEP staff and the participating schools. At the school level, a local administrator was responsible for preparing for and conducting the assessment session in one or more schools. These individuals were usually school or district staff and were trained by Westat staff. In addition, Westat hired and trained a state supervisor for each state. The state supervisors were responsible for working with the state coordinators and overseeing assessment activities. Westat also hired and trained four quality control monitors in each state to monitor 50 percent of the assessment sessions.

5.3 MATERIALS PROCESSING AND DATA ENTRY

Upon completion of each assessment session, field administration personnel shipped the assessment booklets and forms from the field to National Computer Systems for entry into computer files, professional scoring (see section 5.4), checking, and creating the data files for transmittal to ETS. More than 125,000 booklets or questionnaires were received and processed for the Trial State Assessment.

The student data and most of the questionnaire data were transcribed into machine-readable form by scanning the instruments with an optical scanning machine. Data from the school questionnaire were key-entered into the system. An intelligent data entry system was used for resolution of the scanned data, the entry of documents rejected by the scanning machine, and the entry of key-entered information. Additionally, each piece of input data was checked to verify that it was of an acceptable type, that it was within a specified range or ranges of values, and that it was consistent with other data values.

The high volume of collected data and the complexity of the Trial State Assessment processing design, with its spiraled distribution of booklets, as well as the concurrent administration of this assessment and the national assessments, required the development and implementation of flexible, innovative processing programs and a sophisticated process control system. This system allowed an integration of data entry and workflow management systems, including carefully planned and delineated editing, quality control, and auditing procedures.

5.4 PROFESSIONAL SCORING OF MATHEMATICS ITEMS

As assessment materials were received from each school, the student booklets were forwarded to the National Computer Systems professional scoring area to be rated by a team of trained scorers. Like the national assessments, the Trial State Assessment included open-ended items -- items that asked students to provide written responses. Open-ended and multiple-choice items were administered in scannable assessment booklets that were identical to the mathematics booklets used in the eighth-grade national assessment. Scores for the open-ended items in these booklets were gridded in ovals at the bottom of the pages on which the items appeared.

The scoring of the Trial State Assessment was conducted simultaneously with the scoring of the mathematics portion of the national program. The same readers scored the open-ended questions from both programs. The readers for the Trial State Assessment were organized into five teams of twelve readers and one team leader. The five team leaders reviewed discrepancies between readers and reviewed decisions regularly so that all readers scored each item similarly.

5.4.1 Description of Scoring

Each open-ended item had a unique scoring guide that identified the range of possible scores for the item and defined the criteria to be used in evaluating students' responses. Eighteen items were categorized as right/wrong, while 17 items included categories of specific correct and incorrect responses. Various types of incorrect responses were also tracked with separate score points to record information on the specific types of errors students were making.

To obtain statistics on interreader reliability, 20 percent of the responses to each open-ended item (at least 9,200 responses per item) were scored by a second reader. The average interreader reliability over all 35 open-ended items was 97 percent. The reliability was 95 percent or greater for 29 of the 35 items and lower than 90 percent for only two items (86.8 percent and 89.3 percent). The reliability information was used to monitor the capabilities of particular readers and the uniformity across readers for each task.

5.4.2 Open-ended Scores in the Secondary-use Data Files

In the data file codebooks and layouts, open-ended items are identified by "O" in the TYPE field; the range of possible scores and the correct score are given in the layouts in the RANGE and KEY VALUE fields.

5.5 DATABASE CREATION

The data transcription and editing process described in section 5.3 resulted in the transmittal to ETS of four data files: one file for each of the three questionnaires (teacher, school, and excluded student) and one for the student response data. The process of deriving sample weights produced an additional three files of sampling weights which were produced by Westat -- one for students, schools, and excluded students. Before data analyses could be performed, these seven files had to be integrated into a coherent and comprehensive database.

The resulting database consisted of three files for each state -- student, school, and excluded student files. Each record on the student file contained a student's responses to the assessment booklet the student was administered as well as the information from the questionnaire that the student's mathematics teacher completed. Because teacher response data can only be reported at the student level, it was not necessary to have a separate teacher file. The student data file was created by first merging the student response data with the student weights data, then merging the resulting file with the teacher response data. In both steps, the assessment booklet serial number was used as the matching criterion.

The school file was created by merging the school questionnaire file with the school weights file and a file of school variables, supplied by Westat, which included demographic information about the schools collected from the principal's questionnaire. The state and school codes were used as the matching criteria. Since some schools did not return a questionnaire and/or were missing principal's questionnaire data, some of the records in the school file contained only school-identifying information and weights information. Data from the school and student files can be linked through the school code.

The excluded student file was created by merging the excluded student questionnaire file with the excluded student weights file. The assessment booklet serial number was used as the matching criterion.

When the three files -- student, school, and excluded student -- had been created, the database was ready for analysis. Whenever new data values, such as composite background variables or plausible values, were derived, they were added to the appropriate database files using the matching procedures described above.

To evaluate the effectiveness of the quality control of the data entry process, student data from the final integrated database was sampled, and the data were verified in detail against the original instruments received from the field. For this purpose, a number of student booklets were selected at random and compared, character by character, with their representation on the data files. The number of instruments involved in these quality control checks was based on the number needed to establish a statistically reassuring conclusion about the accuracy of the entire data entry operation. Results of the quality control checks are given in *The Technical Report of NAEP's 1990 Trial State Assessment Program.*

Chapter 6

REPORTING SUBGROUPS AND OTHER VARIABLES

Chapter 6: REPORTING SUBGROUPS AND OTHER VARIABLES

6.1 INTRODUCTION

In addition to overall achievement results, the 1990 Trial State Assessment permits reporting on the performance of various subpopulations of the student population. Some reporting subgroups were defined directly from responses to assessment items; some were derived from responses to two or more different items. Section 6.2 defines the reporting subgroups and explains how they are derived.

Certain variables on the data files were formed from the responses to one or more items from the student demographics questionnaire, the student mathematics background questionnaire, the teacher questionnaire, or the school questionnaire. These derived variables are described in section 6.3.

Section 6.4 explains variables that were derived from students' responses to the mathematics items. Section 6.5 provides information about the proficiency variables (the plausible values) and other variables that were used in scaling student response data. Student and school file variables that come from the principal's questionnaire are explained in section 6.6.

Values and response counts for all of the variables described in this chapter are found in the printed codebook for each state. Unless otherwise noted, the variables on the data files are named and defined in the same way for both the state sample and the national winter public-school (NWP) sample that was used for state/nation comparisons.

6.2 REPORTING SUBGROUPS FOR THE 1990 TRIAL STATE ASSESSMENT

Results for the 1990 Trial State Assessment were reported for student subgroups defined by gender, race/ethnicity, type of community, parents' level of education, and geographical region. The following explains how each of these subgroups was derived and the name of the variable to be used to perform secondary analyses of the subgroup data.

DSEX (Gender)

The variable SEX on the student files is the gender of the student being assessed, as taken from school records. For a few students, data for this variable was missing and was imputed by ETS after the assessment. The resulting variable DSEX on the student file contains a value for every student and should be used for gender comparisons among students.

DRACE (Race/ethnicity)

The variable DRACE on the student file is an imputed definition of race/ethnicity, derived from up to three sources of information. This variable is used for race/ethnicity subgroup comparisons. Two items from the student demographics questionnaire were used in the determination of derived race/ethnicity:

Demographic Item Number 2:

2. If you are Hispanic, what is your Hispanic background?

○ I am not Hispanic.
○ Mexican, Mexican American, or Chicano
○ Puerto Rican
○ Cuban
○ Other Spanish or Hispanic background

Students who responded to item number 2 by filling in the second, third, fourth, or fifth oval were considered Hispanic. For students who filled in the first oval, did not respond to the item, or provided information that was illegible or could not be classified, responses to item number 1 were examined in an effort to determine race/ethnicity. Item number 1 read as follows:

Demographic Item Number 1:

1. Which best describes you?

○ White (not Hispanic)

○ Black (not Hispanic)

○ Hispanic ("Hispanic" means someone who is Mexican, Mexican American, Chicano, Puerto Rican, Cuban, or from some other Spanish or Hispanic background.)

○ Asian or Pacific Islander ("Asian or Pacific Islander" means someone who is Chinese, Japanese, Korean, Filipino, Vietnamese, or from some other Asian or Pacific Island background.)

○ American Indian or Alaskan Native ("American Indian or Alaskan Native" means someone who is from one of the American Indian tribes, or one of the original people of Alaska.)

○ Other (What?) ________

Students' race/ethnicity was then assigned to correspond with their selection. For students who filled in the sixth oval ("Other"), provided illegible information or information that could not be classified, or did not respond at all, observed race/ethnicity (RACE on the data files), if provided from school records, was used.

Derived race/ethnicity could not be determined for students who did not respond to background items 1 or 2 and for whom an observed race/ethnicity was not provided.

TOC, SCTOC (Type of community)

NAEP assigned each participating school to one of four type of categories designed to provide information about the communities in which the schools are located. These categories are contained on the student data files as the variable TOC and on the school files as SCTOC.

The type of community categories consist of three "extreme" types of communities and one "other" type of community. Schools were placed into these categories on the basis of information about the type of community, the size of its population (as of the 1980 Census), and an occupational profile of residents provided by school principals before the assessment. The principals completed estimates of the percentage of students whose parents fit into each of six occupational categories. The type of community categories are as follows:

1 - Extreme Rural: Students in this group live outside metropolitan statistical areas, live in areas with a population below 10,000, and attend schools where many of the students' parents are farmers or farm workers.

2 - Disadvantaged Urban: Students in this group live in metropolitan statistical areas and attend schools where a high proportion of the students' parents are on welfare or are not regularly employed.

3 - Advantaged Urban: Students in this group live in metropolitan statistical areas and attend schools where a high proportion of the students' parents are in professional or managerial positions.

4 - Other: Students in this category attend schools in areas other than those defined as advantaged urban, disadvantaged urban, or extreme rural.

PARED (Parents' education level)

The variable PARED on the student file is derived from responses to two questions, B003501 and B003601, in the student demographic questionnaire. Students were asked to indicate the extent of their mother's education (B003501) by choosing one of the following:

○ She did not finish high school.
○ She graduated from high school.
○ She had some education after high school.
○ She graduated from college.
○ I don't know.

Students were asked to provide the same information about the extent of their father's education (B003601) by choosing one of the following:

○ He did not finish high school.
○ He graduated from high school.
○ He had some education after high school.
○ He graduated from college.
○ I don't know.

The information was combined into one parental education reporting category (PARED) as follows: If a student indicated the extent of education for only one parent, that level was included in the data. If a student indicated the extent of education for both parents, the higher of the two levels was included in the data. For students who did not know the level of education for both parents or did not know the level of education for one parent and did not respond for the other, the parental education level was classified as unknown. If the student did not respond for both parents, the student was recorded as having provided no response.

REGION, SREGION (Region of the country)

States were grouped into four geographical regions -- Northeast, Southeast, Central, and West -- as shown in Table 6-1. All 50 states and the District of Columbia are listed, with the participants in the Trial State Assessment highlighted in italic type. Territories were not assigned to a region. Further, the part of Virginia that is included in the Washington, DC, metropolitan statistical area is included in the Northeast region; the remainder of the state is included in the Southeast region.

These data are retained as the variable REGION on the student file and SREGION on the school file.

Table 6-1
NAEP Regions

NORTHEAST	SOUTHEAST	CENTRAL	WEST
Connecticut	***Alabama***	***Illinois***	Alaska
Delaware	***Arkansas***	***Indiana***	***Arizona***
District of Columbia	***Florida***	***Iowa***	***California***
Maine	***Georgia***	Kansas	***Colorado***
Maryland	***Kentucky***	***Michigan***	***Hawaii***
Massachusetts	***Louisiana***	***Minnesota***	***Idaho***
New Hampshire	Mississippi	Missouri	***Montana***
New Jersey	***North Carolina***	***Nebraska***	Nevada
New York	South Carolina	***North Dakota***	***New Mexico***
Pennsylvania	Tennessee	***Ohio***	***Oklahoma***
Rhode Island	***Virginia***	South Dakota	***Oregon***
Vermont	***West Virginia***	***Wisconsin***	***Texas***
Virginia			Utah
			Washington
			Wyoming

Age (DAGE, MODAGE)

Results for students at a particular age can be selected using (1) the student file variable DAGE, the student's age as of December 31, 1989 (i.e., born in 1976), or (2) the student file variable MODAGE. The modal age (the age of most of the students in the grade sample) for the eighth-grade students is age 13. A value of 1 for MODAGE indicates that the student is younger than the modal age; a value of 2 indicates that the student at the modal age; a value of 3 indicates that the student is older than the modal age.

6.3 VARIABLES DERIVED FROM THE STUDENT, TEACHER, AND SCHOOL QUESTIONNAIRES

Several variables on the student data files are formed from the systematic combination of response values for one or more items from either the student demographic questionnaire, the student mathematics background questionnaire, the teacher questionnaire, or the school questionnaire.

These variables maximize use of the data, incorporate a larger segment of the population, and save analysis costs by grouping items that measure similar characteristics into one variable. The derivation of each of these variables is explained below.

HOMEEN2 (Home environment—Articles [of 4] in the home)

The variable HOMEEN2 was created from the responses to student demographic items B000901, B000903, B000904, and B000905 concerning articles found in the student's home (newspaper, encyclopedia, more than 25 books, and magazines). The values for this variable were derived as follows:

1 0-2 types	The student responded to at least two items and answered YES to two or fewer.
2 3 types	The student answered YES to three items.
3 4 types	The student answered YES to four items.
8 Omitted	The student answered fewer than two items.

SINGLEP (How many parents live at home)

SINGLEP was created from items B005601 and B005701, which asked whether the student's mother (or stepmother) and father (or stepfather) lived at home with the student. The values for SINGLEP were derived as follows:

1 2 parents at home	The student answered YES to both items.
2 1 parent at home	The student answered YES to B005601 and NO to B005701, or YES to B005701 and NO to B005601.
3 Neither at home	The student answered NO to both items.
8 Omitted	The student did not respond to or filled in more than one oval for one or both items.

PERCMAT (Students' perception of mathematics)

PERCMAT was created from items M810701 through M810705 in the mathematics background questionnaire, which asked students about their perceptions of each of five statements:

M810701	I like mathematics.
M810702	Almost all people use mathematics in their jobs.
M810703	I am good in mathematics.
M810704	Mathematics is more for boys than for girls.
M810705	Mathematics is useful for solving everyday problems.

For each item, the student could respond as follows:

1. Strongly agree
2. Agree
3. Undecided
4. Disagree
5. Strongly disagree

To derive PERCMAT, first the values for one item (M810704) were reversed (e.g., "strongly disagree" became 1). Then, for each of the five items, values 3, 4, and 5 were combined to create one value (new value 3). PERCMAT was determined by adding the values for the five items and dividing by five to obtain a mean. The mean was then recoded as follows:

1 - 1.67 = 1 Strongly agree
1.68 - 2.33 = 2 Agree
2.34 - 3 = 3 Undecided, disagree, or strongly disagree

The student had to answer at least one of the five items to get a value for PERCMAT.

TCERTIF (Type of teaching certificate)

Items T030301 through T030305 in the teacher questionnaire were combined to produce TCERTIF. The following rules were used to determine the three values for TCERTIF.

1 Mathematics	The teacher responded YES to either T030303 or T030304
2 Education	The teacher responded YES to either T030301 or T030302 and NO to T030303 and T030304
3 Else	Any other response

TUNDMAJ (Undergraduate major)

Items T023301, T023311, T023307, and T023313 in the teacher questionnaire were used to determine TUNDMAJ as follows:

1 Mathematics	The teacher responded YES to T023311
2 Education	The teacher responded YES to T023301 and NO to T023311
3 Else	The teacher responded YES to T023307 or T023313 and NO to T023311 and T023301

TGRDMAJ (Graduate major)

Items T023401, T023411, T023407, and T023413 in the teacher questionnaire were used to determine TUNDMAJ as follows:

1 Mathematics	The teacher responded YES to T023411
2 Education	The teacher responded YES to T023401 and NO to T023411
3 Else	The teacher responded YES to T023407 or T023413 and NO to T023411 and T023401

TMATCRS (Number of mathematics areas in which courses were taken)

TMATCRS was derived from items T030407 through T030411, T030413, and T030414 in the teacher questionnaire. Those items asked how many courses the teacher had taken in a variety of areas. TMATCRS was derived by obtaining a count of the number of times (of seven) that the teacher responded to the number-of-courses category "1," "2," or "3 or more". The levels of TMATCRS were then defined as:

1 0 to 3 courses
2 4 to 5 courses
3 6 to 7 courses

The teacher had to answer at least one of these items to receive a value for TMATCRS.

TEMPHNO (Teacher's emphasis in numbers and operations)

TEMPHNO was derived from teacher questionnaire items T031501, T031502, T031503, T031515, and T031516. The variable was derived by first combining categories three (little emphasis) and four (none) for each item and changing the value for that category to three. The mean of the values for all five items was then recoded as follows:

1 - 1.67	1	Heavy emphasis
1.68 - 2.33	2	Moderate emphasis
2.34 - 3	3	Little or no emphasis

The teacher had to answer at least one of these items to receive a value for TEMPHNO.

TEMPHPS (Teacher's emphasis in data analysis, probability, and statistics)

TEMPHPS was derived from teacher questionnaire items T031506 and T031507. The variable was derived by first combining category three (little emphasis) and four (none) for both items and changing the value for that category to three. The mean of the values for both items was recoded as follows:

1 - 1.67	1	Heavy emphasis
1.68 - 2.33	2	Moderate emphasis
2.34 - 3	3	Little or no emphasis

The teacher had to answer at least one of the items to receive a value for TEMPHPS.

SPOLICY (Changes in school policy since 1984-85)

School questionnaire items C028101 to C028103 and C028105 to C028109 were used to derive the variable SPOLICY. Those items asked if changes had been made in school policy in a variety of areas. SPOLICY was derived by obtaining a count of the number of times (of eight) that the response was YES to these items. The levels of SPOLICY were then defined as:

1 0 to 2 changes
2 3 to 4 changes
3 5 to 8 changes

SPROBS (Problems in the school)

School questionnaire items C028201 through C028211 were used to derive the variable SPROBS. Those items asked if problems existed in the school in a variety of areas. To derive SPROBS, category one (serious) and two (moderate) for each item were combined into a new category one. Category three was recoded as category two and category four was recoded as category three. The mean of the values for all 11 items were then recoded as follows:

1 - 1.67 = 1 Moderate to serious
1.68 - 2.33 = 2 Minor
2.34 - 3 = 3 Not a problem

PCLUNCH (Percent in school lunch program)

The values for the variable PCLUNCH on the student data files were calculated from the school questionnaire variables C025010 (number of students in subsidized lunch program) and C026202 (total enrollment as of October 1, 1989). The value for C025010 was divided by the value for C026202 to create the value for PCLUNCH.

6.4 VARIABLES DERIVED FROM MATHEMATICS ITEMS

CALCUSE (Calculator-usage index)

CALCUSE was created from noncognitive questions included in mathematics blocks M8 and M9. Students were provided a scientific calculator to use in answering the cognitive questions in those two blocks. Each cognitive item was followed by the question *"Did you use a calculator on this question?"*. (These items are identified in the data files and codebooks by the words "(CALC USE)" in the SHORT LABEL field.) The responses to these questions were used to derive the variable CALCUSE.

The cognitive items in blocks M8 (18 items) and M9 (20 items) were classified into one of three categories -- calculator-active, calculator-inactive, and calculator-neutral. Calculator-active items required the use of a calculator for their solution. Calculator-inactive items asked questions for which the use of a calculator was inappropriate. Calculator-neutral items could be solved with or without a calculator. The category for each of the calculator items is identified in column 109 of the machine-readable catalog files for the student data (1 = calculator-active, 2 = calculator-inactive, 3 = calculator-neutral).

Block M8 contained three calculator-active items, seven calculator-inactive items, and eight calculator-neutral items. Block M9 contained five calculator-active items, ten calculator-inactive items, and five calculator-neutral items. Blocks M8 and M9 each appeared in a total of three test booklets. However, one booklet contained both blocks M8 and M9. Therefore, at least one block of calculator items appeared in five of the seven assessment booklets.

The calculator-usage index for students assigned a booklet containing only block M8 was based on 10 items; the index for students assigned a booklet containing only block M9 was based on 15 items; and the index for students assigned a booklet containing both blocks M8 and M9 was based on 25 items.

CALCUSE had two levels, defined as follows:

1 High	Students who used the calculator appropriately (i.e., used it for the calculator-active items and did not use it for the calculator-inactive items) at least 85 percent of the time and indicated they had used the calculator for at least half of the calculator-active items they were presented.
2 Other	Students who did not use the calculator appropriately at least 85 percent of the time or indicated that they had used the calculator for less than half of the calculator-active items they were presented.

The percentage of appropriate calculator usage was determined using only those items that were answered by the student. Omitted items were not included as part of the denominator in calculating the percentage of appropriate calculator use.

NUMCOR (Number correct within booklet)
PCTCOR (Percent correct within booklet)
LOGITP (Logit percent correct within booklet)
ZSCORE (Standardized logit percent correct within booklet)

The student file variables NUMCOR, PCTCOR, and LOGITP are statistics describing a student's responses to the cognitive items in the assessment booklet he or she received. (Note: Each student was administered one of seven different assessment booklets, each of which contained a different combination of mathematics items from the total item pool.) These three variables were used to create a standardized logit score, ZSCORE.

NUMCOR is the number of correct responses a student made to the items in the booklet; PCTCOR is the percent of correct responses, calculated as the number of correct responses (NUMCOR) divided by the total number of items in the booklet. If NUMCOR equaled zero, PCTCOR was set to .0001; if NUMCOR equaled the total number of items in the booklet, PCTCOR was set to .9999.

A logit score, LOGITP, was calculated for each student by the following formula:

$$\text{LOGITP} = \ln\left[\frac{PCTCOR}{1 - PCTCOR}\right]$$

LOGITP was then restricted to a value x, such that $-3 \leq x \leq 3$. After computing LOGITP for each student, the mean and standard deviation was calculated for each booklet as the first step in standardizing the logit scores. The standardized logit score, ZSCORE, was then calculated for each student by the following formula:

$$\text{ZSCORE} = \left[\frac{LOGITP\ -\ mean\ logit}{standard\ deviation}\right]$$

6.5 VARIABLES RELATED TO PROFICIENCY SCALING

Proficiency Score Variables

Item response theory (IRT) was used to estimate average mathematics proficiency for each state and for various subpopulations, based on students' performance on the set of mathematics items they received. IRT provides a common scale on which performance can be reported for the nation, state, and subpopulations, even when all students do not answer the same set of questions. This common scale makes it possible to report on relationships between students' characteristics (based on their responses to the background question) and their overall performance in the assessment.

A scale ranging from 0 to 500 was created to report performance for each of the five mathematics content areas: Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. Each content-area scale was based on the distribution of student performance across all three grades assessed in the 1990 national assessment (grades 4, 8, and 12) and had a mean of 250 and a standard deviation of 50.

Subscale proficiency estimates were obtained for all students assessed in the Trial State Assessment. The NAEP methods use random draws ("plausible values") from estimated proficiency distributions to compute population statistics. Plausible values are not optimal estimates of individual proficiency; instead, they serve as intermediate values to be used in estimating population characteristics. Chapter 7 provides further details on the computation and use of plausible values.

In addition to the subscale plausible values, a composite of the subscales was created as a measure of overall mathematics proficiency. This composite was a weighted average of the subscale plausible values in which the weights were proportional to the relative importance assigned to each content area as specified in the mathematics objectives. The definition of the composite scale for the Trial State Assessment program was identical to that used for the national eighth-grade mathematics program.

The proficiency score (plausible value) variables are provided on the student data files for each of the five subscales and the composite scale and are named as shown in Table 6-2.

Table 6-2
Scaling Variables for the 1990 Trial State Assessment Samples

Mathematics Scale	Data Variables
Numbers and Operations	MRPSCA1 to MRPSCA5
Measurement	MRPSCB1 to MRPSCB5
Geometry	MRPSCC1 to MRPSCC5
Data Analysis, Statistics, and Probability	MRPSCD1 to MRPSCD5
Algebra and Functions	MRPSCE1 to MRPSCE5
Composite	MRPCMP1 to MRPCMP5

SMEANM, SMEANSA (School mean score)
SNSCHM, SNSCHSA (Number of schools ranked)
SRANKM, SRANKSA (School rank)

A mean mathematics composite score (SMEANM on the state school files, SMEANSA on the national comparison sample school files) based on the values from the scaling variable MRPCMP1 was calculated for each school using the students' sampling weights. The schools were then ordered from highest to lowest mean score -- the school with the highest mean score was given a ranking of 1 and the school with the lowest mean score was given a ranking equal to the number of schools in the sample. Values for school rank are found in the variable SRANKM on the state school files and SRANKSA on the national comparison sample school file. The number of schools ranked is found in the variable SNSCHM on the state school files and SNSCHSA on the national comparison sample school files.

These variables were later used in partitioning the schools within the national winter public-school (NWP) comparison sample and the schools within each state into three equal groups based on their ranking (highest third, middle third, and lowest third).

SCHMATH (School-level mathematics mean logit score)

SCHMATH on the student data files is a school-level mean proficiency variable that was used in conditioning procedures (described in Chapter 7) to take into account differences in school proficiency. For each booklet, weighted frequency distributions were obtained (across all states for each state sample and across the full national sample for the NWP comparison sample) of the number of correct responses for the students taking that booklet. A percentile rank for each student was determined from the frequency distribution of the booklet that student received. The logit of the percentile rank was calculated as:

$$\ln\left[\frac{\textit{percentile rank}}{1 - \textit{percentile rank}}\right] .$$

For each school, the weighted mean of the logits for the students in that school was calculated. Each student was then assigned that mean as his or her SCHMATH value.

6.6 PRINCIPAL'S QUESTIONNAIRE VARIABLES (PQ)

Before the assessment, Westat, Inc., distributed a questionnaire to the principal of each participating school to gather data about school characteristics, including parents' occupations and student race/ethnicity. The data variables from this questionnaire are retained on the school file. A subset of these variables are also on the student files. Principal's questionnaire variables are identified in the data layouts by "(PQ)" in the SHORT LABEL field.

Chapter 7

NAEP SCALING PROCEDURES AND THEIR APPLICATION IN THE TRIAL STATE ASSESSMENT

Chapter 7: NAEP SCALING PROCEDURES AND THEIR APPLICATION IN THE TRIAL STATE ASSESSMENT

7.1 INTRODUCTION

The primary method by which results from the Trial State Assessment are disseminated is scale-score reporting. With scaling methods, the performance of a sample of students in a subject area or subarea can be summarized on a single scale or series of subscales even when different students have been administered different items. Sections 7.2 through 7.5 present an overview of the scaling methodologies employed in the analyses of the data from NAEP surveys in general and from the Trial State Assessment in particular. Details of the scaling procedures specific to the Trial State Assessment are presented in section 7.6.

7.2 THEORETICAL BACKGROUND OF NAEP SCALING PROCEDURES

The basic information from an assessment consists of the responses of students to the items presented in the assessment. For NAEP, these items are generated to measure performance on sets of objectives developed by nationally representative panels of learning area specialists, educators, and concerned citizens. Satisfying the objectives of the assessment and ensuring that the tasks selected to measure each goal cover a range of difficulty levels typically requires a large number of items. The Trial State Assessment required 137 items. To reduce student burden, each assessed student was presented only a fraction of the full pool of items using multiple matrix sampling procedures.

The most direct manner of presenting the assessment results is to report percent correct statistics for each item. However, because of the vast amount of information, separate results for each of the items in the assessment pool hinders the comparison of the general performance of subgroups of the population. Item-by-item reporting ignores overarching similarities in trends and subgroup comparisons that are common across items.

It is useful to view the assessed items as random representatives of a conceptually infinite pool of items within the same domain and of the same type. In this random item concept, a set of items is taken to represent the domain of interest. An obvious measure of achievement within a domain of interest is the average percent correct across all presented items within that domain. The advantage of averaging is that it tends to cancel out the effects of peculiarities in items which can affect item difficulty in unpredictable ways. Furthermore, averaging makes it possible to compare more easily the general performances of subpopulations.

Despite their advantages, there are a number of significant problems with average percent correct scores. First, the interpretation of these results depends on the selection of the items; the selection of easy or difficult items could make student performance appear to be overly high or low. Second, the average percent correct metric is related to the particular items comprising the average, so that direct comparisons in performance between subpopulations require that those subpopulations have been administered the same set of items. Third, because this approach limits comparisons to percents correct on specific sets of items, it provides no simple way to report trends over time when the item pool changes. Finally, the average percent correct provides no estimate of the distribution of proficiency in the population when each student is administered only a fraction of the items. Average percent correct statistics describe the mean performance of students within subpopulations but provide no other information about the distributions of skills among students in the subpopulations.

These limitations can be overcome by the use of response scaling methods. If several items require similar skills, the regularities observed in response patterns can often be exploited to characterize both respondents and items in terms of a relatively small number of variables. When combined through appropriate mathematical formulas, these variables capture the dominant features of the data. Furthermore, all students can be placed on a common scale, even though none of the respondents take all of the items within the pool. Using the scale, it becomes possible to discuss distributions of proficiency in a population or subpopulation and to estimate the relationships between proficiency and background variables.

It is important to point out that any procedure of aggregation, from a simple average to a complex multidimensional scaling model, highlights certain patterns at the expense of other potentially interesting patterns that may reside within the data. Every item in a NAEP survey is of interest and can provide useful information about what young Americans know and can do. The choice of an aggregation procedure must be driven by a conception of just which patterns are salient for a particular purpose.

The scaling for the Trial State Assessment was carried out within the five mathematics content areas specified in the objectives because it was anticipated that different patterns of performance might exist for these essential subdivisions of the subject area. The five subscales corresponded to one of the following content areas: Numbers and Operations; Measurement; Geometry; Data Analysis, Statistics, and Probability; and Algebra and Functions. By creating a separate subscale for each of these content areas, potential differences in subpopulation performance between the content areas are maintained. Analyses of the subscale level results from the 1990 Trial State Assessment and national mathematics assessment have shown that the subscales provide additional information that a single scale cannot -- for example gender differences in mathematics performance by subscale.

The creation of subscales to describe mathematics performance does not preclude the reporting of an overall mathematics composite as a single index of overall mathematics performance. A composite is computed as the weighted average of the subscale scores where the weights correspond to the relative importance given to each subscale as defined by the objectives. The composite scores provide a global measure of performance within the subject area while the constituent subscale scores allow the measurement of important interactions within educationally relevant subdivisions of the subject area.

7.3 SCALING METHODOLOGY

This section reviews the scaling models employed in the analyses of data from the Trial State Assessment and the 1990 national mathematics assessment, as well as the "plausible

BEST COPY AVAILABLE

values" methodology that allows such models to be used with NAEP's sparse item-sampling design. The reader is referred to Mislevy (in press) for an introduction to plausible values methods and a comparison with standard psychometric analyses, to Mislevy and Sheehan (1987) and Beaton and Johnson (1990) for additional information on how the models are used in NAEP, and to Rubin (1987) for the theoretical underpinnings of the approach.

The 137 mathematics items administered in the Trial State Assessment were also administered to eighth-grade students in the national mathematics assessment. However, because the administration procedures differed, the Trial State Assessment data was scaled independently from the national data. The national data also included results for fourth- and twelfth-grade students. Details of the scaling of the Trial State Assessment and the subsequent linking to the results from the national mathematics assessment are provided in section 7.6.

7.3.1 The Scaling Model

The scaling model used by NAEP in the Trial State Assessment is the three-parameter logistic (3PL) model from item response theory (IRT; e.g., Lord, 1980). This is a "latent variable" model, defined separately for each of the five subscales, and quantifying respondents' tendencies to provide correct answers to the items contributing to a subscale as a function of a parameter that is not directly observed, called proficiency on the subscale.

The fundamental equation of the 3PL model is the probability that a person whose proficiency on subscale k is characterized by the *unobservable* variable θ_k will respond correctly to item j:

$$\begin{aligned} P(x_j=1 \mid \theta_k, a_j, b_j, c_j) &= c_j + (1-c_j)/\{1+\exp[-1.7a_j(\theta_k - b_j)]\} \\ &\equiv P_j(\theta_k), \end{aligned} \tag{7.1}$$

where

x_j	is the response to item j, 1 if correct and 0 if not;
a_j	where $a_j > 0$, is the slope parameter of item j, characterizing its sensitivity to proficiency;
b_j	is the threshold parameter of item j, characterizing its difficulty; and
c_j	where $0 \leq c_j < 1$, is the lower asymptote parameter of item j, reflecting the chances of a correct response from students of very low proficiency; c parameters are estimated for multiple-choice items, but are fixed at zero for open-ended items.

A typical assumption of item response theory is the conditional independence of the probabilities of correct response by an individual to a set of items, given the individual's proficiency. That is, conditional on the individual's θ_k, the joint probability of a particular

BEST COPY AVAILABLE

response pattern $\underline{x} = (x_1,\ldots,x_n)$ across a set of n items is simply the product of terms based on equation (7.1):

$$P(\underline{x}|\theta_k,\underline{a},\underline{b},\underline{c}) = \prod_j^n [P_j(\theta_k)]^{x_j}[1 - P_j(\theta_k)]^{1-x_j} \tag{7.2}$$

It is also typically assumed that response probabilities are conditionally independent of background variables ($\underline{y}$), given θ_k, or

$$P(\underline{x}|\theta_k,\underline{a},\underline{b},\underline{c},y) = p(\underline{x}|\theta_k,\underline{a},\underline{b},\underline{c}). \tag{7.3}$$

After $\underline{x}$ has been observed, equation (7.2) can be viewed as a likelihood function, and provides a basis for inference about θ_k or about item parameters. Estimates of item parameters were obtained with a modified version of Mislevy and Bock's (1982) BILOG computer program, then treated as known in subsequent calculations. The parameters of the items constituting each of the five subscales were estimated independently of the parameters of the other subscales. Once items have been calibrated in this manner, a likelihood function for the subscale proficiency θ_k is induced by a vector of responses to any subset of calibrated items, thus allowing θ_k-based inferences from matrix samples.

As stated previously, item parameter estimation was performed independently for the Trial State Assessment and for the national mathematics assessment. In both cases, the identical subscale definitions were used. The national mathematics data also included responses of fourth-grade students to 109 mathematics items and responses of twelfth-grade students to 144 mathematics items, where 45 items were common between grades 4 and 8 and 63 items were common between grades 8 and 12. The subscales for national mathematics extends across the three grades.

Conditional independence is a mathematical assumption, not a necessary fact of nature. Although the IRT models are employed in NAEP only to summarize average performance, a number of checks are made to detect serious violations of conditional independence, and, when warranted, remedial efforts are made to mitigate its effects on inferences. These checks include the following:

1) Checks on relative item operating characteristics among distinct gender and ethnicity groups (i.e., differential item functioning, or DIF [Holland & Thayer, 1988]). Some degree of relative differences are to be expected, of course, and modestly varying profiles among groups will exist beyond the differences conveyed by their differing θ distributions. The intent of the check at this stage is to detect and eliminate items that operate differentially for identifiable reasons that are unrelated to the skills intended to be measured in the subject area.

2) When a subscale extends over age groups as is the case for the national mathematics subscales, evidence is sought of different operating characteristics over ages. When such effects are found, an item in question is represented by different item parameters in different age groups.

Item-level factor analyses have diminished in importance as our perspective of the role of IRT in NAEP has evolved. The assumption that performance in a scaling area is driven by a single unidimensional variable is unarguably incorrect in detail. However, our use of the model is not theoretical, instead it is data analytic; interpretation of results is not trait-referenced, but domain-referenced. Scaling areas are determined *a priori* by considerations of content as collections of items for which overall performance is deemed to be of interest. The IRT summary is not expected to capture all meaningful variation in item response data, but to reflect distributions of overall proficiency -- to summarize the main patterns in item percents-correct in the populations and subpopulations of interest. Using a unidimensional IRT model when the true model is multidimensional captures these overall patterns even though it over- or under-estimates the covariances among responses to items in pairs. For inferences based on overall proficiency, violations of the model with respect to dimensionality are less serious than violations in the shapes of the marginal response curves -- hence our greater attention to routine checks of item-fit residuals for every item in every calibration run than to factor analytic results.

In all NAEP IRT analyses, missing responses at the end of each block a student was administered were considered "not-reached," and treated as if they had not been presented to the respondent. Missing responses before the last observed response in a block were considered intentional omissions, and treated as fractionally correct at the value of the reciprocal of the number of response alternatives. These conventions are discussed by Mislevy and Wu (1988). With regard to the handling of not-reached items, Mislevy and Wu found that ignoring not-reached items introduces slight biases into item parameter estimation to the degree that not-reached items are present and speed is correlated with ability. With regard to omissions, they found that the method described above provides consistent limited-information likelihood estimates of item and ability parameters under the assumption that respondents omit only if they can do no better than responding randomly.

The local independence assumption embodied in equation (7.2) implies that item response probabilities depend only on θ and the specified item parameters--not on the position of the item in the booklet, on the content of items around an item of interest, or on test-administration timing conditions. These effects are certainly present in any application. The practical question is whether the IRT probabilities obtained via (7.2) are "close enough" to be robust with respect to the context in which the data are to be collected and the inferences that are to be drawn.

The experience with adaptive testing has shown using the same item parameters regardless of when an item is administered does not materially bias estimates of the proficiencies of individual examinees. Our experience with the 1986 NAEP reading anomaly, has shown, however, that for measuring small changes over time, changes in item context and speededness conditions lead to unacceptably large random error components. These can be avoided by presenting items used to measure change in identical test forms, with identical timings and administration conditions. Thus we do *not* maintain that the item parameter estimates obtained in any particular booklet configuration are appropriate for other conceivable configurations, and the parameter estimates are context-bound. (For this reason, we prefer common population equating to common item equating whenever equivalent random samples are available for linking.) This is the reason that the data from the Trial State Assessment were calibrated separately from the data from the national NAEP -- since the administration procedures differed somewhat between the Trial State Assessment and the national NAEP, the values of the item parameters could be different.

7.3.2 An Overview of Plausible Values Methodology

Item response theory was developed in the context of measuring individual examinees' abilities. In that setting, each individual is administered enough items (often 100 or more) to permit precise estimation of his or her θ, as a maximum likelihood estimate $\hat{\theta}$, for example. Because the uncertainty associated with each θ is negligible, the distribution of θ, or the joint distribution of θ with other variables, can then be approximated using individuals' $\hat{\theta}$ values as if they were θ values.

This approach breaks down in the assessment setting when, in order to provide broader content coverage in limited testing time, each respondent is administered relatively few items in a scaling area. The problem is that the uncertainty associated with individual θs is too large to ignore, and the features of the $\hat{\theta}$ distribution can be seriously biased as estimates of the θ distribution. (The failure of this approach was verified in early analyses of the 1984 NAEP reading survey; see Wingersky, Kaplan, & Beaton, 1987.) "Plausible values" were developed as a way to estimate key population features consistently, and approximate others no worse than standard IRT procedures would. A detailed development of plausible values methodology is given in Mislevy (in press). Along with theoretical justifications, that paper presents comparisons with standard procedures, discussions of biases that arise in some secondary analyses, and numerical examples.

The following provides a brief overview of the plausible values approach, focusing on its implementation in the Trial State Assessment analyses.

Let $\underline{y}$ represent the responses of all sampled examinees to background and attitude questions, along with design variables such as school membership, and let $\underline{\theta}$ represent the subscale proficiency values. If $\underline{\theta}$ were known for all sampled examinees, it would be possible to compute a statistic $t(\underline{\theta},\underline{y})$ -- such as a subscale or composite subpopulation sample mean, a sample percentile point, or a sample regression coefficient -- to estimate a corresponding population quantity T. A function $U(\underline{\theta},\underline{y})$ -- e.g., a jackknife estimate -- would be used to gauge sampling uncertainty, as the variance of t around T in repeated samples from the population.

Because the 3PL model is a latent variable model, however, $\underline{\theta}$ values are not observed even for sampled students. To overcome this problem, we follow Rubin (1987) by considering $\underline{\theta}$ as "missing data" and approximate $t(\underline{\theta},\underline{y})$ by its expectation given $(\underline{x},\underline{y})$, the data that actually were observed, as follows:

$$\begin{aligned} t^*(\underline{x},\underline{y}) &= E[t(\underline{\theta},\underline{y})\,|\,\underline{x},\underline{y}] \\ &= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}\,|\,\underline{x},\underline{y})\, d\underline{\theta} \ . \qquad (7.4) \end{aligned}$$

It is possible to approximate t^* using random draws from the conditional distributions, $p(\underline{\theta}\,|\,x_i,y_i)$, of the subscale proficiencies given the item responses x_i and background variables y_i

for sampled student i. These values are referred to as "imputations" in the sampling literature, and "plausible values" in NAEP. The value of $\underline{\theta}$ for any respondent that would enter into the computation of t is thus replaced by a randomly selected value from the conditional distribution $p(\underline{\theta}|x_i,y_i)$. Rubin (1987) proposes that this process be carried out several times--"multiple imputations" -- so that the uncertainty associated with imputation can be quantified. The average of the results of, for example, M estimates of t, each computed from a different set of plausible values, is a Monte Carlo approximation of (7.4); the variance among them, B, reflects uncertainty due to not observing θ, and must be added to the estimated expectation of $U(\underline{\theta},\underline{y})$, which reflects uncertainty due to testing only a sample of students from the population. Section 7.3 explains how plausible values are used in subsequent analyses.

It cannot be emphasized too strongly that **plausible values are *not* test scores for *individuals*** in the usual sense. Plausible values are offered only as intermediary computations for calculating integrals of the form of equation (7.4), in order to estimate *population* characteristics. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated. The key idea lies in a contrast between plausible values and the more familiar θ estimates of educational measurement that are in some sense optimal for each examinee (e.g., maximum likelihood estimates, which are consistent estimates of an examinee's θ, and Bayes estimates, which provide minimum mean-squared errors with respect to a reference population): *Point estimates that are optimal for individual examinees have distributions that can produce decidedly nonoptimal (specifically, inconsistent) estimates of population characteristics* (Little & Rubin, 1983). Plausible values, on the other hand, are constructed explicitly to provide consistent estimates of population effects.

7.3.3 Computing Plausible Values in IRT-based Scales

Plausible values for each respondent i are drawn from the conditional distribution $p(\theta|x_i,y_i)$. This subsection describes how, in IRT-based scales, these conditional distributions are characterized, and how the draws are taken. An application of Bayes' theorem with the IRT assumption of conditional independence produces

$$\begin{aligned} p(\underline{\theta}|x_i,y_i) &\propto P(x_i|\underline{\theta},y_i)\, p(\underline{\theta}|y_i) \\ &= P(x_i|\underline{\theta})\, p(\underline{\theta}|y_i), \end{aligned} \qquad (7.5)$$

where, for vector-valued $\underline{\theta}$, $P(x_i|\underline{\theta})$ is the product over subscales of the *independent likelihoods* induced by responses to items within each subscale, and $p(\underline{\theta}|y_i)$ is the multivariate--and generally nonindependent -- *joint density* of proficiencies for the subscales, conditional on the observed value y_i of background responses.

In the analyses of the data from the Trial State Assessment and the data from the national mathematics assessment, a normal (Gaussian) form was assumed for $p(\underline{\theta}|y_i)$, with a common dispersion and with a mean given by a linear model based on the first 90 - 95 principal

BEST COPY AVAILABLE

components of 170 selected main-effects and two-way interactions of the complete vector of background variables. The included background variables will be referred to as the *conditioning variables*, and will be denoted y^c. (The conditioning variables used in the Trial State Assessment analyses are listed in *The Technical Report of NAEP's 1990 Trial State Assessment*.) The following model was fit to the data within each state:

$$\theta = \Gamma' y^c + \varepsilon , \tag{7.6}$$

where ε is normally distributed with mean zero and dispersion Σ. The number of principal components of the conditioning variables used for each state was sufficient to account for 90 percent of the total variance of the full set of conditioning variables (after standardizing each variable). As in regression analysis, Γ is a matrix each of whose columns is the *effects* for one subscale and Σ is the matrix *variance of residuals* between subscales. By fitting the model (7.6) separately within each state, interactions between each state and the conditioning variables are automatically included in the conditional joint density of subscale proficiencies. Like item parameter estimates, the estimates of the parameters of conditional distributions were treated as known true values in subsequent steps of the analyses.

Maximum likelihood estimates of Γ and Σ were obtained with Sheehan's (1985) M-GROUP computer program, using a variant of the EM solution described in Mislevy (1985). The difference from the published algorithm lies in the numerical approximation that was employed. Note from (7.5) that $p(\underline{\theta}|x_i,y_i)$ is proportional to the product of two terms, the likelihood $P(x_i|\underline{\theta})$ and the conditional distribution $p(\underline{\theta}|y_i)$. The conditional distribution for person i has been assumed multivariate normal, with mean $\mu_i^c = \Gamma' y_i^c$ and covariance matrix Σ; if the likelihood is approximated by another normal distribution, with mean μ_i^L and covariance matrix Σ_i^L, then the posterior $p(\underline{\theta}|x_i,y_i)$ is also multivariate normal with covariance matrix

$$\Sigma_i^P = (\Sigma^{-1} + (\Sigma_i^L)^{-1})^{-1} \tag{7.7}$$

and mean vector

$$\bar{\underline{\theta}}_i = (\underline{\theta}_i^c \Sigma^{-1} + \underline{\theta}_i^L (\Sigma_i^L)^{-1})(\Sigma_i^P)^{-1}. \tag{7.8}$$

In the analyses of the Trial State Assessment, a normal approximation for $P(x_i|\underline{\theta})$ is accomplished in a given scale by the steps described below. (Recall that by the assumed conditional independence across scales, the joint conditional likelihood for multiple scales is the product of independent likelihoods for each of the scales.) These computations are carried out in the scale determined by BILOG (Mislevy & Bock, 1982) item parameter estimates, where the

BEST COPY AVAILABLE

mean and standard deviation of the composite population formed by combining the three NAEP grade/ages has mean zero and standard deviation one. The steps were as follows.

1) Lay out a grid of Q equally spaced points from -5 to +5, a range that covers the region in each scale where all examinees are virtually certain to occur. The value of Q varies from 20 to 40, depending on the subscale being used; smaller values suffice for subscales with few items given to each respondent, while larger values are required for subscales with many items.

2) At each point X_q, compute the likelihood $L(x_i|\theta=X_q)$.

3) To improve the normal approximation in those cases in which likelihoods are not approximately symmetric in the range of interest -- as when all of a respondent's answers are correct -- multiply the values from Step 2 by the mild smoothing function

$$S(X_q) = \frac{\exp(X_q+5)}{[1+\exp(X_q+5)][1+\exp(X_q-5)]}. \qquad (7.9)$$

   This is equivalent to augmenting each examinee's response vector with responses to two fictitious items, one extraordinarily easy item that everyone gets right and one extraordinarily difficult item that everyone gets wrong. This expedient improves the normal approximation for examinees with flat or degenerate likelihoods in the range where their conditional distributions lie, but has negligible effects for examinees with even modestly well-determined symmetric likelihoods.

4) Compute the mean and standard deviation of θ using the weights $S(X_q)L(x_i|\theta=X_q)$ obtained in Step 3.

At this stage the likelihood induced by a respondent's answers to the items in a given scale is approximated by a normal distribution. Since the mathematics assessment uses five subscales, independent normal distributions, one per subscale, are used to summarize information from responses to items from the several subscales.

This normalized-likelihood/normal posterior approximation was then employed in both the estimation of Γ and Σ and in the generation of plausible values. From the final estimates of Γ and Σ, a respondent's posterior distribution was obtained from the normal approximation using the four-step procedure outlined above. A plausible value was drawn from this multivariate normal distribution. Finally, weighted-average composites over subscales were also calculated after appropriate rescaling.

7.4 ANALYSES

When survey variables are observed without error from every respondent, standard variance estimators quantify the uncertainty associated with sample statistics from the only source of the uncertainty, namely the sampling of respondents. Item percents correct for NAEP cognitive items meet this requirement, but scale-score proficiency values do not. The IRT models used in their construction posit an unobservable proficiency variable θ to summarize performance on the items in the subarea. The fact that θ values are not observed even for the respondents in the sample requires additional statistical analyses to draw inferences about θ distributions and to quantify the uncertainty associated with those inferences. As described above, Rubin's (1987) multiple imputations procedures were adapted to the context of latent variable models to produce the plausible values upon which many analyses of the data from the Trial State Assessment were based. This section describes how plausible values were employed in subsequent analyses to yield inferences about population and subpopulation distributions of proficiencies.

7.4.1 Computational Procedures

Even though one does not observe the θ value of respondent i, one does observe variables that are related to it: x_i, the respondent's answers to the cognitive items he or she was administered in the area of interest, and y_i, the respondent's answers to demographic and background variables. Suppose one wishes to draw inferences about a number $T(\underline{\theta},\underline{Y})$ that could be calculated explicitly if the θ and y values of each member of the population were known. Suppose further that if θ values were observable, we would be able to estimate T from a sample of N pairs of θ and y values by the statistic $t(\underline{\theta},\underline{y})$ [where $(\underline{\theta},\underline{y}) = (\theta_1,y_1,\ldots,\theta_N,y_N)$], and that we could estimate the variance in t around T due to sampling respondents by the function $U(\underline{\theta},\underline{y})$. Given that observations consist of (x_i,y_i) rather than $(\underline{\theta}_i,y_i)$, we can approximate t by its expected value conditional on $(\underline{x},\underline{y})$, or

$$\begin{aligned} t^*(\underline{x},\underline{y}) &= E[t(\underline{\theta},\underline{y})\,|\,\underline{x},\underline{y}] \\ &= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta}\,|\,\underline{x},\underline{y})\, d\underline{\theta}\,. \qquad (7.10) \end{aligned}$$

It is possible to approximate t^* with random draws from the conditional distributions $p(\underline{\theta}_i|x_i,y_i)$, which are obtained for all respondents by the method described in section 7.3.3. Let $\underline{\theta}_m$ be the m^{th} such vector of "plausible values," consisting of a multidimensional value for the latent variable of each respondent. This vector is a plausible representation of what the true $\underline{\theta}$ vector might have been, had we been able to observe it.

The following steps describe how an estimate of a scalar statistic $t(\underline{\theta},\underline{y})$ and its sampling variance can be obtained from M (>1) such sets of plausible values. (Five sets of plausible values are used in NAEP analyses of the Trial State Assessment.)

1) Using each set of plausible values $\hat{\underline{\theta}}_m$ in turn, evaluate t as if the plausible values were true values of $\underline{\theta}$. Denote the results $\hat{t}_m$, for m=1,...,M.

2) Using the jackknife variance estimator defined in Chapter 8, compute the estimated sampling variance of $\hat{t}_m$, denoting the result U_m.

3) The final estimate of t is

$$t^* = \sum_{m=1}^{M} \frac{\hat{t}_m}{M}. \tag{7.11}$$

4) Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents:

$$U^* = \sum_{m=1}^{M} \frac{U_m}{M}. \tag{7.12}$$

5) Compute the variance among the M estimates $\hat{t}_m$, to approximate uncertainty due to not observing θ values from respondents:

$$B_M = \sum_{m=1}^{M} \frac{(\hat{t}_m - t^*)^2}{(M - 1)} \tag{7.13}$$

6) The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1})\, B_M. \tag{7.14}$$

Note: Due to the excessive computation that would be required, NAEP analyses did not compute and average jackknife variances over all five sets of plausible values, but only on the first set. Thus, in NAEP reports, U^* is approximated by U_1.

7.4.2 Statistical Tests

Suppose that if θ values were observed for sampled students, the statistic $(t - T)/U^{1/2}$ would follow a t-distribution with d degrees of freedom. Then the incomplete-data statistic $(t^* - T)/V^{1/2}$ is approximately t-distributed, with degrees of freedom given by

$$\nu = \frac{1}{\dfrac{f_M^2}{M - 1} + \dfrac{(1 - f_M)^2}{d}} \tag{7.15}$$

where f_M is the proportion of total variance due to not observing θ values:

$$f_M = (1+M^{-1})\, B_M / V_M \ . \tag{7.16}$$

When B_M is small relative to U^*, the reference distribution for incomplete-data statistics differs little from the reference distribution for the corresponding complete-data statistics. This is the case with main NAEP reporting variables. If, in addition, d is large, the normal approximation can be used to flag "significant" results.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U_m and U^* is a covariance matrix, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity

$$(T-t^*)\, V^{-1}\, (T-t^*)' \tag{7.17}$$

is approximately F distributed, with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of f_M:

$$f_M = (1+M^{-1}) \text{ Trace } (B_M V_M^{-1})/k. \tag{7.18}$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices.

7.4.3 Biases in Secondary Analyses

Statistics t^* that involve proficiencies in a scaled content area and variables included in the conditioning variables y^c, are consistent estimates of the corresponding population values T. Statistics involving background variables y that were *not* conditioned on, or relationships among proficiencies from *different* content areas, are subject to asymptotic biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variables to the variables that were conditioned on and to the proficiency of interest. That is, the large sample expectations of certain sample statistics need not equal the true population parameters.

The *direction* of the bias is typically to underestimate the effect of nonconditioned variables. For details and derivations see Beaton and Johnson (1990), Mislevy (in press), and Mislevy and Sheehan (1987, section 10.3.5). For a given statistic t^* involving one content area and one or more nonconditioned background variables, the *magnitude* of the bias is related to the extent to which observed responses x account for the latent variable θ, and the degree to which the nonconditioned background variables are explained by conditioning background variables. The first factor -- conceptually related to test reliability -- acts consistently in that greater measurement precision reduces biases in *all* secondary analyses. The second factor acts to reduce biases in certain analyses but increase it in others. In particular,

- High shared variance between conditioned and nonconditioned background variables *mitigates* biases in analyses that involve only proficiency and nonconditioned variables, such as marginal means or regressions.

- High shared variance *exacerbates* biases in regression coefficients of conditional effects for nonconditioned variables, when nonconditioned and conditioned background variables are analyzed jointly as in multiple regression.

The large number of background variables that have been included in the conditioning vector for the Trial State Assessment allows a large number of secondary analyses to be carried out with little or no bias, and mitigates biases in analyses of the marginal distributions of θ in nonconditioned variables. Kaplan and Nelson's analysis of the 1988 NAEP reading data (some results of which are summarized in Mislevy, in press), which had a similar design and fewer conditioning variables, indicate that the potential bias for nonconditioned variables in multiple regression analyses is below 10 percent, and biases in simple regression of such variables is below 5 percent. Additional research (summarized in Mislevy, in press) indicates that most of the bias reduction obtainable from conditioning on a large number of variables can be captured by instead conditioning on the first several principal components of the matrix of all original conditioning variables. This procedure was adopted for the Trial State Assessment by replacing the 170 conditioning effects by the first K principal components, where K was selected so that 90 percent of the total variance of the full set of conditioning variables (after standardization) was captured. Mislevy (in press) shows that this puts an upper bound of 10 percent on the potential bias for all analyses involving the original conditioning variables.

7.5 SCALE ANCHORING

Scale anchoring is a method for attaching meaning to a scale. Traditionally, meaning has been attached to educational scales by norm-referencing -- that is, by comparing students at a particular scale level to other students. In contrast, the NAEP scale anchoring is accomplished by describing what students at selected levels know and can do. This is the primary purpose of NAEP.

The anchoring process was performed on the national NAEP mathematics composite as follows. Composite plausible values for each student (in grades 4, 8, and 12 and/or for age 9, 13, and 17) who participated in the national mathematics assessment were created as a weighted average of the subscale plausible values, where each set of plausible values for a particular subscale was linearly adjusted to have a mean of 250.5 and a standard deviation of 50. The scale levels 200, 250, 300 and 350 on the 500 point scale were selected. These values (roughly standard deviation units apart) are far enough apart to be noticeably different but not so far apart as to be trivial.

The students are sorted by their plausible values, and students with a plausible value at or near each level (i.e. within 12.5 points) are grouped together. For the group at the lowest scale score level, what they know and can do is defined by the items that at least 65 percent of the students answered correctly. At a higher score level, the question is: what is it that students at this level know and can do that students at the next lower level cannot. The answer is defined by the items that at least 65 percent of students at this level answered correctly, but a majority (at least 50 percent) at the next lower level answered incorrectly. Finally, the difference between the probabilities of success between the two levels must be at least 30 percentage points. The assessment items are, therefore, grouped by the levels between which they discriminate. It is important to note that the overall percentage of students who correctly answer an anchor item is not equal to the percentage scoring above that scale level.

Table 7-1 demonstrates the statistical anchoring process. Three items are displayed, identified by the labels "A", "B", and "C". Four anchoring levels are identified, corresponding to scale values of 200, 250, 300 and 350. In the table, Item "A" anchors at the 250 level since the probability of correct response for students with proficiencies around 250 is 80 percent while the probability of success for students at the next lower level (200) is 40 percent. Item "B" anchors at the 300 level since there is a steep rise in the probability of success between 250 and 300 and since the probabilities of success at the two levels satisfy the threshold values. Item "C" does not anchor at any of the four levels because the discrimination between adjacent levels is not sufficiently sharp. Of the 275 unique items in the 1990 national mathematics assessment, 143 (52 percent) satisfied the anchoring criteria with an additional 53 (19 percent) nearly satisfying the criteria.

Table 7-1
Three Example Items for Scale Anchoring:
Percentages of Students Scoring at or near the Scale Value
Who Responded Correctly to the Item

Item	Scale Values			
	200	250	300	350
A	40%	80%	87%	92%
B	20%	23%	68%	84%
C	30%	56%	81%	87%

Following the determination of the anchor levels, a committee of mathematics experts, educators, and others was assembled to review the items and, using their knowledge of mathematics and student performance, to generalize from the items to more general constructs. To derive the descriptions of the four scale anchor points, the 19 panelists first worked in two independent groups and then as a whole. Although the two sets of descriptions did not differ substantively, the group felt that the cross-validation procedure was valuable. As a final step, the reconciled version was sent to all panelists for review.

7.6 SCALING THE 1990 TRIAL STATE ASSESSMENT DATA

This section describes some of the details of the analyses carried out in developing the Trial State Assessment content area scales and composite scale. The philosophical and theoretical underpinnings of the NAEP scaling procedures were described in the previous sections of this chapter.

The first step in the analysis of the Trial State Assessment data involved conventional item and test analyses -- for example, examinations of average proportions correct and average biserial correlations. These analyses are discussed in detail in *The Technical Report of NAEP's 1990 Trial State Assessment*. This section focuses on the four major steps in the *scaling* of the Trial State Assessment data:

- Item response theory (IRT) scaling
- Estimation of state and subgroup proficiency distributions based on the "plausible values" methodology
- Linking of the Trial State Assessment content area scales to the corresponding scales from the 1990 national assessment
- Creation of the Trial State Assessment mathematics composite scale

An overview of each of these steps is provided in the following sections. The rationale for and details of the steps are given in *The Technical Report of NAEP's 1990 Trial State Assessment*.

7.6.1 Item Response Theory (IRT) Scaling

IRT-based content area scales were developed, using the 3-parameter logistic (3PL) model described in section 7.3, by separately calibrating the sets of items in each of the five content areas. Item parameter estimates on a provisional scale were obtained using a modified version of the BILOG program (Mislevy & Bock, 1982). The BILOG item calibrations were based on the data from a systematic random sample of about 25 percent of the students who participated in the Trial State Assessment. This sample of students (650 students from each of the 40 participating jurisdictions) will be referred to as the "calibration sample."

The Trial State Assessment analysis plans called for a single set of item parameters to be estimated for each item. This common set of item parameters was to be used for obtaining the scaled score results for all 40 states and for both monitored and unmonitored sessions. To obtain a single set of item parameters in which 1) sampling weights were used to reflect the demographic composition within each state, 2) each state's data contributed equally to the estimation process, and 3) data from monitored and unmonitored sessions contributed equally, the final sampling weights[1] were rescaled only for item parameter estimation.

---

[1] The weights provided by Westat for estimating total-group and subgroup statistics.

BEST COPY AVAILABLE

The sampling for item calibration and the rescaling of weights included the following:

- Samples of 650 records were drawn for each state using systematic sampling -- 325 from the monitored sessions and 325 from the unmonitored sessions. This resulted in a total sample of 26,000 records.

- For each state, the sum of the Westat sampling weights for the set of monitored and unmonitored records selected for the sample was obtained (these sums are denoted as WM_s and WU_s, respectively).

- For each state, the Westat weights for the individuals in the sample (denoted as w_{si}) were rescaled so the sum of the weights for the monitored and unmonitored sessions would each be equal to 325. Thus, for the monitored students in the sample,

$$w^*_{si} = w_{si}\,(325/WM_s\,),$$

and for the unmonitored students,

$$w^*_{si} = w_{si}\,(325/WU_s\,),$$

where w^*_{si} denotes the rescaled weight for individual i from state s.

IRT calibrations were carried out separately for each scale using a version of the BILOG program which has been modified for use in NAEP. Prior distributions were imposed on item parameters with the following starting values: thresholds (normal[0,2]); slopes (log-normal[0,.5]); and, asymptotes (2-parameter beta with parameter values determined as functions of the number of response options for an item and a weight factor of 50). The locations (but, not the dispersions) were updated at each program estimation cycle in accordance with provisional estimates of the item parameters. Items presented to, but not reached by, students were treated as "not-presented" items. Intentional omissions were treated as fractionally correct with probability equal to the reciprocal of the number of response options for each item.[2]

Item parameter estimation proceeded in two phases. First, the subject ability distribution was assumed fixed [normal(0,1)] and a stable solution was obtained. The parameter estimates from this solution were then used as starting values for a subsequent set of runs in which the subject ability distribution was freed and estimated concurrently with item parameter estimates. After each estimation cycle, the subject ability distribution was restandardized to have a mean of zero and standard deviation of one and, correspondingly, parameter estimates for that cycle were also linearly restandardized.

Model fit was evaluated by examining BILOG likelihood ratio chi-square statistics[3] and by examining plots of nonmodel-based estimates of the expected conditional (on θ) proportion

---

[2] The probability was set at zero for open-ended items.

[3] These sampling distributions of these statistics are probably not strictly χ^2 with the indicated degrees of freedom. Therefore, they were used as descriptive indices of relative model fit rather than in a statistically rigorous fashion.

BEST COPY AVAILABLE

correct versus the proportion correct predicted by the estimated ICC at each of set of θ levels (see Mislevy & Sheehan, 1987, p. 302). In general, the fit of the model was quite good. During the estimation process, difficulties obtaining a stable set of parameter estimates were encountered for only two of the 137 items. One of these items was removed from the measurement scale since preliminary graphical analyses suggested a poor fit to the 3PL model. This item also was removed from the Measurement scale for the national assessment. For the other item (algebra and functions scale), the 3PL model appeared to fit well, however, difficulty was encountered obtaining a converged slope estimate. A decision was made to retain the item and fix the slope at the value obtained after ten BILOG estimation cycles. The IRT parameters for the items included in the Trial State Assessment are listed in Appendix B.

7.6.2 Estimation of State and Subgroup Proficiency Distributions

The proficiency distributions (for the total population in each state, and for important subgroups within each state) were estimated by using the multivariate plausible values methodology described in the previous sections (see also Mislevy, 1988). The background variables included in the model (denoted $\underline{y}$ in section 7.3) were principal component scores derived from the correlation matrix of selected main-effects and two-way interactions associated with a wide range of student, teacher, school, and community variables. A set of five multivariate plausible values was drawn for each individual who participated in the Trial State Assessment.

Plans for reporting each jurisdiction's results required analyses examining the relationships between proficiencies and a large number of background variables. The background variables included student demographic characteristics (e.g., the race/ethnicity of the student, highest level of education attained by parents), student attitudes toward mathematics, student behaviors both in and out of school (e.g., amount of television watched daily, amount of mathematics homework each day), the type of mathematics class being taken (e.g., algebra, or general eighth-grade mathematics), the amount of emphasis on various topics included in the assessment provided by the students' teachers, as well as a variety of other aspects of the students' background and preparation, the background and preparation of their teachers, and the educational, social, and financial environment of the schools they attended. Overall, relationships between proficiency and more than 50 variables, taken directly or derived from the student, teacher, and school questionnaires, or provided by Westat, were estimated and reported.

To avoid biases in reporting results and to minimize biases in secondary analyses, it is desirable to incorporate measures of all variables to be reported on as independent variables in the conditioning model. When expressed in terms of contrast-coded main effects and interactions, the variables cited above resulted in 167 variables to be included in the conditioning model. (A listing of the complete set of variables included in the conditioning model is provided in *The Technical Report of NAEP's 1990 Trial State Assessment.*)

BEST COPY AVAILABLE

The conditioning model included up to 167 contrasts[4]. Many of these contrasts were highly correlated with other contrasts in the model; other contrasts involved relatively small numbers of individuals. Under such conditions, it can be difficult to obtain converged estimates of Γ and Σ (described in the previous sections) based on the iterative numerical procedures used in the computer program M-GROUP (Sheehan, 1985), which is used by NAEP to estimate conditioning models and generate plausible values. To minimize such potential convergence problems, the original background variable contrasts were standardized and transformed into a set of linearly independent variables by extracting principal components from the correlation matrix of the original contrast variables. The principal components, rather than the original variables, were used as the independent variables in the conditioning model.

Principal components are a set of uncorrelated linear combinations of the original standardized variables (Harris, 1975). They retain information about variability and intercorrelation among the original variables. Previous analyses of the NAEP 1988 Reading data suggested that conditioning using principal components virtually eliminated biases in analyses involving the original effects from which the components were derived (Mislevy, 1988). In addition, because principal components are uncorrelated, the M-GROUP estimation problems which might have resulted from the high degree of multicollinearity among the original variables were avoided.

The same variables and codings were included in the conditioning model for all 40 Trial State Assessment participants. (These variables and codings are given in *The Technical Report of NAEP's 1990 Trial State Assessment.*) In addition, a single common set of IRT item parameters (shown in Appendix B of this guide) were used. However, principal components were extracted separately and separate conditioning models were estimated for each of the 40 Trial State Assessment participants.

In theory, the number of principal components that could be extracted is equal to the total number of the original contrast variables minus the number of these variables that are exactly collinear with other variables (or collections of variables) in the model. Analyses by Kaplan and Nelson (see Mislevy, in press) on the 1988 NAEP reading data suggested that a relatively small number of principal components will capture almost all of the variance and most of the complex intercorrelations among the set of original variables and will reduce most of the potential bias for primary and secondary analyses. For the Trial State Assessment analysis, the number of principal components included for each state was that number required to account for approximately 90 percent of the variance in the original contrast variables.

7.6.3 Linking State and National Scales

One of the purposes of the Trial State Assessment Program was to allow each participating jurisdiction to compare its results with the nation as a whole and with the region of the country in which that jurisdiction is located.[5] To permit such comparisons, a nationally

---

[4] In some states, one or more contrasts were not possible since all individuals were at the same level of that contrast.

[5] No regions are designated for the territories.

BEST COPY AVAILABLE

representative sample of public-school students in the eighth-grade was tested as part of the national assessment using the same assessment booklets as in the Trial State Assessment. In addition, a subsample of the national assessment was tested at about the same time of the year (January to March 1990) as were students participating in the Trial State Assessment (February 5 to March 2, 1990).

For valid comparisons to be made between each of the Trial State Assessment participants and the relevant national subsample, results from the two assessments had to be expressed in terms of a similar system of scale units. As described above, the provisional BILOG scales for the Trial State Assessment (and subsequent estimation of proficiency distributions using plausible values) were computed independently from the scaling used for the national assessment[6].

A procedure analogous to linearly equating test forms was used to link the Trial State Assessment and national scales. The Trial State Assessment and national scales were made comparable in the sense that estimated proficiency distributions from two samples (the Trial State Assessment and a special sample of the national assessment [called the State Aggregate Comparison Sample and described below]) from the same population (eighth-grade students in public schools in the 37 states and the District of Columbia) were constrained to have the same mean and standard deviation.[7]

The State Aggregate Comparison (SAC) sample was a subsample of 2,467 students from the winter subsample of the national assessment. The SAC subsample consists of all eighth-grade students in public schools in the 37 participating states and the District of Columbia who were assessed as part of the winter administration of the national mathematics assessment. With appropriate weighting (provided by Westat), the SAC is a representative sample of the population of all grade-eligible public-school students within the 37 states and the District of Columbia participating in the Trial State Assessment and was assessed at a reasonably similar point in time as the Trial State Assessment.

The following steps were followed to linearly link the scales of the two assessments:

1) For each scale, an estimate of the proficiency distribution of the total Trial State Assessment sample (minus the students from Guam and the Virgin Islands) was obtained using the full set of plausible values generated by the M-GROUP program. Recall that these plausible values are expressed on the provisional Trial State Assessment scale and were generated using the common state item parameters, but separate state-specific conditioning coefficients. The weights used were the final sampling weights. Thus, the resulting estimate pertains to the distribution of

---

[6] Care was taken to ensure that the five scales were produced for both the national and Trial State Assessment, and that all the items included in the Trial State Assessment were also included in the national assessment. Because the national assessment spans three age/grades, additional items are used in developing the national scales which were not part of the Trial State Assessment.

[7] Data from the two territories (Guam and the Virgin Islands) were excluded for the purposes of establishing the link to the national scale.

BEST COPY AVAILABLE

proficiency in the aggregated group of eighth-grade public-school students in the 37 states and the District of Columbia.

The arithmetic mean of the five sets of plausible values was taken as the estimated mean of the Trial State Assessment distribution, and the geometric mean of the standard deviations of the five sets of plausible values was taken as the estimated standard deviation of the distributions for each scale.

2) For each scale, an estimate of the proficiency distribution of the total SAC subsample of the national eighth-grade winter half-sample was obtained using the full set of plausible values for this group. These plausible values were expressed in terms of the scale that was intended to be used for reporting the results for the national mathematics assessment and were generated using the national assessment item parameters and a common set of eighth-grade specific conditioning coefficients. The weights used were specially provided by Westat to allow for the estimation of proficiency for the same population of students as for state data (i.e., the aggregated group of eighth-grade public-school students in 37 states and the District of Columbia).

   The means and standard deviations of the distributions for each scale were obtained for this sample in the same manner as described in step 1.

3) For each content area scale, a set of linear transformation coefficients to link the state scale to the corresponding national scale were obtained. The linking was of the form,

$$Y^* = \alpha + \beta X_{TSA}$$

   where,

   X_{TSA} = a scale level in terms of the system of units of the provisional BILOG scale

   Y^* = scale level in terms of the system of units comparable to those used for reporting the national mathematics results

   β = (SD_{SAC}/SD_{TSA}),

   α = $(M_{SAC} - \beta M_{TSA})$

   SD_{SAC} = the estimated standard deviation of the SAC sample proficiency distribution

   SD_{TSA} = the estimated standard deviation of the Trial State Assessment equating sample proficiency distribution (with Guam and Virgin Islands removed)

BEST COPY AVAILABLE

M_{SAC} = the estimated mean of the SAC sample proficiency distribution

M_{TSA} = the estimated mean of the Trial State Assessment equating sample proficiency distribution (with Guam and Virgin Islands removed)

7.6.4 Producing a Mathematics Composite Scale

For the national assessment, a grade 8 composite scale was created as an overall measure of mathematics proficiency for students at that grade. The composite was a weighted average of plausible values on the five content area scales (Numbers and Operations; Measurement; Geometry; Data Analysis, Probability, and Statistics; and Algebra and Functions). The weights for the national scale were proportional to the relative importance assigned to each content area in the assessment specifications developed by the Mathematics Objectives Panel. The weights for each content area were similar to the actual proportion of items from that content area in the entire eighth-grade item pool.

A Trial State Assessment composite scale was developed using weights identical to those used to produce the grade 8 composite for the 1990 national mathematics assessment. The weights were as follows:

Table 7-2
Weights for the Composite Scale

Content Area Scale	Weight for Composite	Proportion of Item Pool
Numbers and Operations	.30	.34
Measurement	.15	.15
Geometry	.20	.19
Data Analysis, Probability, and Statistics	.15	.14
Algebra and Functions	.20	.18

In developing the Trial State Assessment composite, the weights were applied to the plausible values for each content area scale as expressed in terms of the final Trial State Assessment scales (i.e., after transformation from the provisional BILOG scales.)

7.6.5 Proficiency Means for the 1990 Trial State Assessment Mathematics Scales

Table 7-3 shows the average mathematics proficiencies for students in the national winter public-school (state/nation comparison) sample. Average proficiencies are given for each subscale and the composite scale for each of the five plausible values and their mean. A similar table for each state is included at the beginning of each state's data codebooks.

BEST COPY AVAILABLE

Table 7-3

Average Mathematics Proficiencies by Scale and Plausible Value for the 1990 National Winter Public-School Comparison Sample

Scale	Data Variables	1st Value	2nd Value	3rd Value	4th Value	5th Value	Mean Value (s.e.)*
Numbers and Operations	MRPSCA1-5	266.04	265.04	265.80	265.14	265.92	265.59 (1.45)
Measurement	MRPSCB1-5	257.01	257.26	257.91	257.30	258.45	257.59 (1.74)
Geometry	MRPSCC1-5	258.75	259.29	259.47	259.09	259.07	259.13 (1.45)
Data Analysis, Statistics, and Probability	MRPSCD1-5	261.61	261.87	262.54	261.22	261.94	261.84 (1.76)
Algebra and Functions	MRPSCE1-5	259.83	260.27	260.47	260.47	260.07	260.22 (1.29)
Composite	MRPCMP1-5	261.32	261.29	261.80	261.23	261.66	261.46 (1.40)

---

* The standard error is the square root of two variance components: the estimated sample variance and the variance due to measurement error.

Chapter 8

CONDUCTING STATISTICAL ANALYSES OF 1990 NAEP TRIAL STATE ASSESSMENT DATA

Chapter 8: CONDUCTING STATISTICAL ANALYSES OF 1990 NAEP TRIAL STATE ASSESSMENT DATA

8.1 INTRODUCTION

Standard statistical procedures should not be applied to the NAEP Trial State Assessment data without modification because the special properties of the data affect the validity of conventional techniques of statistical inference. There are two reasons for this. First, a complex sampling scheme, rather than simple random sampling, was used to collect NAEP data. Second, because scaling models were used to summarize performance in each subject area, measurement error must be taken into account when analyzing scale-score proficiency variables.

In the NAEP sampling scheme, students do not have an equal probability of being selected. Therefore, as in all complex surveys, each student has been assigned a sampling weight. The larger the probability of selection for students within a particular demographic group, the smaller the weights for those students will be. When computing descriptive statistics or conducting inferential procedures, one should weight the data for each student. ***Performance of statistical analyses without weights can lead to misleading results.***

Another way in which the complex sample design used by NAEP differs from simple random sampling is that the NAEP sampling scheme involves the selection of clusters of students from the same school, as well as clusters of schools from urbanicity, income, and minority strata (in the case of the Trial State Assessment) and from the same geographically defined primary sampling unit, or PSU (in the case of the national assessment). As a result, observations are not independent of one another as they are in a simple random sample. Therefore, ***use of standard formulas for estimating the standard error of sample statistics such as means, proportions, or regression coefficients will result in values that are generally too small.*** The standard error, which is a measure of the variability of a sample statistic, gives an indication of how well that statistic estimates the corresponding population value. It is used to conduct tests of statistical significance. If conventional simple random sampling formulas are used to compute standard errors, too many statistically significant results will occur in most instances.

Another effect of the NAEP sampling scheme is a reduction of the effective degrees of freedom. In a simple random sample, the degrees of freedom of a variance estimate are based primarily on the number of subjects (although it also depends on the distribution of the variable under consideration). In the NAEP 1990 designs, the degrees of freedom are a function of the number of clusters of schools (for the Trial State Assessment) or clusters of PSUs (for the national assessment) and the number of strata in the design, rather than the number of subjects (see Chapter 4 for a discussion of the sample design). Therefore, ***the standard formulas for obtaining degrees of freedom are not valid with the NAEP data.***

Proficiencies in mathematics content areas were summarized through item response theory (IRT) scaling models, but not in the way that these models are used in standard

BEST COPY AVAILABLE

applications in which enough responses are available from each person to estimate his or her proficiency precisely. NAEP administers relatively few items to each respondent in order to track *population* levels of proficiency more efficiently. Because the data are not intended to estimate *individual* levels of proficiency, however, more complicated analyses are required.

The following sections outline the procedures used in NAEP to account for the special properties of the data. Section 8.2 discusses the use of weights to account for the differential sampling rates and certain other adjustments, such as for nonresponse. Section 8.3 discusses jackknife procedures that can be used to estimate sampling variability. Section 8.4 describes the "plausible values" that can be used to estimate population levels of proficiency in the subject areas, and shows how to use them in analyses. Section 8.5 suggests simpler approximations for the procedures described in 8.3 and 8.4, such as using design effects rather than the jackknife to estimate sampling variability. Although this procedure is less precise, it requires substantially less computation. We expect that the resulting degree of accuracy will be acceptable to most users of NAEP data.

8.2 USING WEIGHTS TO ACCOUNT FOR DIFFERENTIAL REPRESENTATION

The 1990 Trial State and national assessments used complex sample designs to obtain the students who were assessed. The goal of the national design was to obtain a series of samples (for the various ages and grades) from which estimates of population and subpopulation characteristics could be obtained with reasonably high precision (low sampling variability) per unit of cost. The goal of the Trial State design was to obtain a sample of students for each state from which estimates of population and subpopulation characteristics could be obtained with approximately equal precision for all states.

To accomplish these goals, NAEP used multistage cluster sample designs (described in Chapter 4) in which the probabilities of selection of the clusters were proportional to measures of their size. To provide improved precision in the estimation of the characteristics of various subpopulations of interest, in the national assessment some subpopulations (corresponding to students from areas with high concentrations of Black or Hispanic students) were deliberately sampled at approximately twice the normal rate to obtain larger samples of respondents from those subpopulations. The result of these differential probabilities of selection for the national assessment is a series of achieved samples, each containing proportionately more members of certain subgroups than there are in the population.

Appropriate estimation of population characteristics for both the Trial State Assessment and national assessment samples must take the sampling design into account. This is accomplished by assigning a weight to each respondent, where the weights properly account for the sample design and, in the case of the national assessment, reflect the appropriate proportional representation of the various types of individuals in the population. These weights also include adjustments for nonresponse and, in the case of the national assessment, adjustments (known as poststratification adjustments) designed to make sample estimates of certain subpopulation totals conform to external, more accurate, estimates. An overview of the computation of these weights appears in Chapter 4. For the present purpose, it is sufficient to note that these weights should be used for all analyses, whether exploratory or confirmatory.

The 1990 Trial State Assessment database includes a number of different samples from several populations. Each of these samples has its own set of weights to be used to produce estimates about the characteristics of the population addressed by the sample (the target population). The various samples, their target populations, and their weights are discussed in the following sections.

8.2.1 The 1990 State Samples of Students

These samples, one for each state, consist of all students assessed in that state as part of the Trial State Assessment. The target population for each state consists of all eighth-grade students enrolled in public secondary schools who were deemed assessable by their school. Either of two alternatively scaled weights can be used for analyses at the student level. The first of these, ORIGWT, has been scaled so that the sum of weights for all students in each state estimates the total number of assessable eighth-grade students in that state's secondary public schools. The second of these, WEIGHT, is a proportional rescaling of ORIGWT, carried out so that the sum of WEIGHT across students and states is equal to the total Trial State Assessment sample size across all states (i.e., the total number of assessed students in the Trial State Assessment). Both weights should provide identical estimates of means, proportions, correlations, and other statistics of interest in analyses within each state as well as analyses involving data from more than one state.

An estimate of the proportion of students in the population who possess some characteristic can be obtained using either WEIGHT or ORIGWT as the ratio of the sum of the weights for the students with that characteristic, divided by the sum of the weights for all students sampled from that population. In the case where ORIGWT is used, the numerator of the proportion is the estimated total number of students with that characteristic and the denominator is the estimated population total. Estimated proportions can also be restricted to subpopulations. For example, the estimated proportion of all assessable students from advantaged urban schools in New York is

$$\frac{WTOT(\textit{New York and Advantaged Urban})}{WTOT(\textit{New York})}$$

where WTOT(New York and Advantaged Urban) is the sum of the weights (WEIGHT or ORIGWT) of all students in New York who are in advantaged urban schools and WTOT(New York) is the sum of the weights (WEIGHT or ORIGWT) of all students in New York.

It is also clearly of interest to estimate the relative proportion of a population (say New York students) who could correctly respond to an assessment exercise. This proportion is estimated by the ratio

$$P = \frac{WTOT(\textit{New York, answered item correctly})}{WTOT(\textit{New York, presented the item})}$$

where the numerator is the sum of weights (WEIGHT or ORIGWT) of all assessed students in New York who responded to the item correctly and the denominator is the sum of weights (WEIGHT or ORIGWT) of all students who

1) were from New York, and,
2) were presented the item (i.e., reached the item, including those who reached it and left it blank).

This total is less than WTOT(New York) because not all students are presented every item, either as a result of the spiral design or as a result of not reaching the item. However, the sample of assessed students in New York who had an opportunity to respond to the item (which includes those who did not reach the item) is itself a representative sample of the entire population of assessable students in New York.

8.2.2 Special Trial State Assessment Comparison Weights for Monitored and Unmonitored Sessions

Within each state, a random half of all assessment sessions were observed by Westat quality control monitors. Investigators may be interested in assessing the impact of monitoring on assessment performance or in otherwise comparing the samples of students in the monitored and unmonitored sessions. For example, it might be of interest to compare the percentage of students from monitored sessions in New York that correctly answered a particular mathematics question to the corresponding percentage from unmonitored sessions. For such analyses, special comparison weights have been provided. As with the overall weight, two alternative scalings are available, CWEIGHT (which sums to the overall sample size) and CORIGWT (which sums to population sizes). Either CWEIGHT or CORIGWT should be used in lieu of WEIGHT or ORIGWT for all analyses intended to compare statistics (such as a mean, proportion, or correlation) obtained from monitored sessions to the same statistic obtained in the unmonitored sessions.

8.2.3 The Winter Public-school Sample from the National Assessment

One of the purposes of the Trial State Assessment was to allow each participating state to compare its results with the nation as a whole, and with the region of the country in which that state is located. To permit such comparisons, a nationally representative sample of students was tested as part of the national assessment using the same assessment booklets as were students participating in the Trial State Assessment. There were, however, some differences between the Trial State Assessment and the full national assessment sample. The Trial State Assessment samples were restricted to public-school students in the eighth grade, while the national sample included public- and private-school students who either were attending eighth grade or were 13 years old. In addition, the entire Trial State Assessment sample was tested in the winter of 1990 (February) while only half of the national assessment sample was tested at a comparable time. The other half-sample of the national assessment was tested in the spring of 1990.

In order to allow for valid state/nation comparisons, a national winter public-school (NWP) sample was created from the full national assessment sample and is included on the Trial State Assessment data files. The NWP sample consists of students from the only the winter half-sample of the national assessment and includes only eighth-grade students enrolled in public schools. As with the Trial State Assessment samples, two sets of weights are available for use with the NWP sample. ORIGWT will sum to the size of the NWP population. WEIGHT is a proportional rescaling of ORIGWT whose sum is approximately equal to the NWP sample size. When used with standard statistical packages, both sets of weights will produce identical results for point estimates of means, proportions, standard deviations, correlations, and other such statistics.

8.2.4 School-based Weights

The 1990 Trial State and national assessments collected questionnaire data from the assessed students' teachers about their background and instructional practices and information from administrators about aspects of the schools attended by the assessed students. Analyses of these data using the weights described above will produce results that are focused on students (e.g., *What percentage of students attend schools in which mathematics is taught by teachers with bachelor's degrees in mathematics?*). For the school questionnaire data, it possible to conduct school-level analyses (e.g., *In what proportion of schools do teachers with bachelor's degrees in mathematics teach mathematics classes?*). The school weights SCHWTF should be used for these purposes. It should be noted that analogous teacher weights are not provided and the NAEP samples were not selected to contain representative samples of teachers. Analyses of the teacher questionnaire data should be restricted to student-level analyses.

8.3 PROCEDURES USED BY NAEP TO ESTIMATE SAMPLING VARIABILITY (Jackknifing)

This section describes how the sampling variability of statistics based on the NAEP data can be estimated. The jackknife variance estimator described below gives fairly precise estimates of the total sampling error for population estimates derived from NAEP student and school data, and for conducting multivariate analyses. To aid secondary users who have fewer resources than those available for the NAEP reports, section 8.5 provides a less expensive approximation for estimating sampling variances.

A major source of uncertainty in the estimation of the value in the population of a variable of interest exists because information about the variable is obtained on only a sample from the population. To reflect this fact, it is important to attach to any statistic (e.g., a mean) an estimate of the sampling variability to be expected for that statistic.

Estimates of sampling variability provide information about how much the value of a given statistic would be likely to change if the statistic had been based on another equivalent sample of individuals drawn in exactly the same manner as the achieved sample. Consequently, the estimation of the sampling variability of any statistic must take into account the sample design.

The NAEP samples are obtained via a stratified multistage probability sampling design that includes, in the case of the NWP, provisions for sampling certain subpopulations at higher rates. Additional characteristics of the sample include adjustments for both nonresponse and, for the NWP, poststratification. The resulting sample has different statistical characteristics than those of a simple random sample. In particular, because of the effects of cluster selection (students within schools, and for the NWP, schools within PSUs) and nonresponse and other weighting adjustments, observations made on different students cannot be assumed to be independent of each other. Furthermore, to account for the differential probabilities of selection and the various sample weighting adjustments, each student has an associated sampling weight that must be used in the computation of any statistic and is itself subject to sampling variability.

Treatment of the data as a simple random sample, with disregard for the special characteristics of the NAEP sample design, will produce underestimates of the true sampling variability. A procedure known as jackknifing is suitable for estimating sampling errors from such a complex design. This procedure has a number of properties that make it particularly suited to the analysis of NAEP data:

1) It provides unbiased estimates of the sampling error arising from the complex sample selection procedure for linear estimates such as simple totals and means, and does so approximately for more complex estimates.

2) It reflects the component of sampling error introduced by the use of weighting factors, such as nonresponse adjustments, that are dependent on the sample data actually obtained.

3) It can be adapted readily to the estimation of sampling errors for parameters estimated using statistical modeling procedures, as well as for tabulation estimates such as totals and means.

4) Once appropriate weights are derived and attached to each record, jackknifing can be used to estimate sampling errors. A single set of replicate weights is required for all tabulations and model parameter estimates that may be needed.

Here the method of applying the jackknife procedure involves first defining pairs (or occasionally triples) of replicate groups. For the Trial State Assessment, a replicate group consists of a school, two or three schools, or (for the largest schools selected with certainty) random subgroups of students within schools. For the national assessments, a replicate group consisted of a single PSU, a pair of PSUs, or (for the large certainty PSUs) schools within a PSU. The replicate groups were paired in accordance with the sample design. The pairing is done independent of performance information obtained from the sample. For the 1990 assessment, Westat defined 56 such pairs for both the national assessment and the Trial State Assessment. These pairings are identified by the variable REPGRP on the NWP student data files and REPGRP1 and REPGRP2 on the state student data files; membership within the pair (or triple) is identified by the variable DROPWT on the NWP student data files and DROPGRP on the state student data files (on the school files, these names are preceded with "S").

Components of the sampling variability of an estimate are each estimated as the squared difference between the value of the statistic for the complete sample and a pseudoreplicate formed by recomputing the statistic on a specially constructed pseudodataset. This pseudodataset is created from the original dataset by eliminating one member of a pair and replacing it with a copy of the remaining unit or units in the pair. For computational purposes, a pseudoreplicate associated with a given pair is the original dataset with a different set of weights (referred to as the student replicate weights SRWT01 through SRWT56 on the data files, where SRWTi is for the i^{th} pair). This set of weights allows measurement of the total effect of replacing one member of the pair with a copy of the other(s), including adjustments for nonresponse and, for the NWP, poststratification. The i^{th} pseudoreplicate for a given statistic is obtained by recalculating the statistic using the weights SRWTi instead of the original sampling weights.

As a specific example of the use of the student replicate weights, let $t(\underline{y},\underline{w})$ be any statistic that is a function of the sample responses $\underline{y}$ and the weights $\underline{w}$ that estimates population value T. For example, t could be a weighted mean, a weighted percent-correct point, or a weighted regression coefficient. The $t(\underline{y},\underline{w})$, computed with the sampling weights (WEIGHT on the data files) is the appropriate sample estimate of T. To estimate $\hat{Var}(t)$, the sampling variance for this statistic, proceed in the following manner:

1) For each of the 56 pairs of first-stage units, compute the associated pseudoreplicate for the statistic. For the i^{th} pair, this is

$$t_i = t(\underline{y}, \underline{SRWTi}) \; ,$$

which is the statistic t recalculated by using SRWTi instead of the original sampling weights.

2) The estimated sample variance of t is

$$\hat{Var}(t) = \sum_{i=1}^{56} (t_i - t)^2 \; .$$

We refer to this estimation technique as the multiweight jackknife approach. Tables 10-7 and 10-8 in Chapter 10 provide SPSS-X and SAS code for carrying out the above in the special case of a weighted mean.

Replicate weights have been provided for:

1) Overall analyses in each state in the Trial State Assessment samples	SRWT01 to SRWT56
2) For monitored/unmonitored comparisons within each state in the Trial State Assessment sample	CSRWT01 to CSRWT56

3)	For school-based analyses in each state for the Trial State Assessment samples	SCHWT01 to SCHWT56
4)	Overall analyses in the NWP sample	SRWT01 to SRWT56
5)	For school-based analyses in the NWP sample	SCHWT01 to SCHWT56

In addition, for analyses comparing national and state results, or for comparisons among states, an appropriate single set of replicate weights can be formed for the merged dataset by using the relevant set of replicate weights for each given component. That is, the first replicate estimate of a difference between a national student-level estimate and that for a given state is obtained by using the replicate weight SRWT01 for each record in the national sample and for each record in the particular state sample, and calculating the difference between the respective replicated national and state estimates.

As a very simple example of how the jackknife variance estimate is computed, consider the following cut-down example, designed to demonstrate the steps. Although the full set of NAEP data consists of thousands of observations and 56 student replicate weights, for the example we will consider a dataset with eight observations and two student replicate weights. Furthermore, the weights have been simplified for clarity.

Table 8-1
Example Dataset to Demonstrate the Jackknife

First-stage Unit	REPGRP	DROPWT	Y	WEIGHT	SRWT01	SRWT02
1	1	1	5	10	20	10
1	1	1	4	9	18	9
2	1	2	6	12	0	12
2	1	2	3	8	0	8
3	2	1	8	4	4	8
3	2	1	9	6	6	12
4	2	2	7	5	5	0
4	2	2	10	4	4	0

In the example dataset there are four first-stage units, 1 through 4, each consisting of two of the eight observations. The first-stage units are divided into two pairs, as identified by the column REPGRP. Within each of those pairs, one first-stage unit is designated as the first member of the pair (REPGRP = 1) while the other is designated as the second (REPGRP = 2). The statistic of interest is the weighted average of the response Y using the weights WEIGHT, and is equal to

$$t = NUM/DEN = 5.914$$

where

$$NUM = 10\times5 + 9\times4 + 12\times6 + 8\times3 + 4\times8 + 6\times9 + 5\times7 + 4\times10 = 343$$

is the weighted sum of the responses and

$$DEN = 10 + 9 + 12 + 8 + 4 + 6 + 5 + 4 = 58$$

is the sum of the weights WEIGHT.

The first pseudoreplicate of the statistic t is the weighted mean recomputed using the SRWT01 as the weights and is

$$t_1 = NUM_1/DEN_1 = 5.842$$

where

$$NUM_1 = 20\times5 + 18\times4 + 0\times6 + 0\times3 + 4\times8 + 6\times9 + 5\times7 + 4\times10 = 333$$

and

$$DEN_1 = 20 + 18 + 0 + 0 + 4 + 6 + 5 + 4 = 57\ .$$

Similarly, $t_2 = 354/59 = 6$ is the weighted mean computed using SRWT02 as the weights. The jackknife variance estimate is then

$$\begin{aligned} V\hat{a}r(t) &= (t_1 - t)^2 + (t_2 - t)^2 \\ &= (-0.072)^2 + (0.086)^2 = .01258 \end{aligned}$$

and the jackknife standard error of t is .112, the square root of the variance.

8.3.1 Degrees of Freedom of the Jackknifed Variance Estimate

The effective number of degrees of freedom of the variance estimate $V\hat{a}r(t)$ will be at most equal to the number of pairs used in forming the pseudoreplicates. The number of degrees of freedom in sampling from normally distributed variates with uniform variances is sufficient information to indicate the variability of the variance estimate, and is equal to the number of independent pieces of information used to generate the variance. For the main assessment sample, the pieces of information are the 56 squared differences $(t_i - t)^2$, each supplying at most one degree of freedom, regardless of how many individuals were sampled within any PSU. (There are fewer pairs with the bridge samples and the season-specific samples and, consequently, fewer degrees of freedom.)

The effective number of degrees of freedom of the sample variance estimator can be less than the number of pairs (56) if the differences are not normally distributed or if some of the squared differences $(t_i - t)^2$ are markedly different in magnitude than others. An extreme case of the latter is when one or more of the t_i are identical to t, so that $(t_i - t)^2 = 0$. This may happen, for example, when the statistic t is a mean for a subgroup, such as a geographic region, and no members of that subgroup come from the pair i. Such a pair contributes zero to the effective number of degrees of freedom of the variance estimate.

An estimate of the effective number of degrees of freedom for $\hat{Var}(t)$ comes from an approximation due to Satterthwaite (1946). (See Cochran, 1977, p. 96, for a discussion.)

If the t_i are normally distributed, the effective number of degrees of freedom using this approximation is

$$df_{eff} = \frac{\left[\sum_{i=1}^{m} (t_i - t)^2 \right]^2}{\sum_{i=1}^{m} (t_i - t)^4} ,$$

where M is the number of pairs used (for the Trial State Assessment, m = 56).

Empirical evidence indicates that this approximation greatly underestimates the degrees of freedom for the sum of single-degree-of-freedom chi-square random variables. Johnson and Rust (in press) propose the following adjustment to df_{eff}:

$$df_{adj} = (3.18 - \frac{2.8}{m^{.5}})\ df_{eff}$$

This adjustment, derived from the results of a simulation study, returns the correct degrees of freedom (m) for the sum of m independent one-degree-of-freedon chi-square random variables to within +/- .6 for $5 \leq m \leq 100$.

8.3.2 Estimation of Subpopulations with Appropriate Jackknifed Standard Errors

As stated in section 8.2.1, WEIGHT, CWEIGHT, XWEIGHT, and SCHWTF are proportional rescalings of, respectively, ORIGWT, CORIGWT, XORIGWT, and SORIGWT. Table 8-2 gives the factors used to calculate the rescaled weight from the original for each Trial State Assessment sample.

These factors are required to estimate the number in a population and compute the corresponding jackknifed standard error, which estimates how well the number in the population has been estimated. The replicate weights SRWT01 to SRWT56 are on the WEIGHT rescaling metric. To use the jackknife procedure with ORIGWT, multiply each replicate weight by the appropriate factor, yielding new replicate weights to be used in the jackknife procedure. The resulting standard error will be the appropriate estimate of the variability of the weights.

Table 8-2
Factors Used to Calculate the Rescaled Weight (WEIGHT) from the Original Weight (ORIGWT)

Sample	Original Weight	Factor		Rescaled Weight
		State	National (NWP)	
Student Overall	ORIGWT	22.343	952.69	WEIGHT
Student Comparison (Monitored vs. Unmonitored)	CORIGWT	44.686	------	CWEIGHT
Excluded Student	XORIGWT	24.801	------	XWEIGHT
School	SORIGWT	4.957	253.88	SCHWTF

8.4 PROCEDURES USED BY NAEP TO HANDLE MEASUREMENT ERROR

Jackknifing provides a reasonable estimate of uncertainty due to the sampling of respondents when the variable of interest is observed without error from every respondent. Population percents correct for cognitive items meet this requirement, but scale-score proficiency values do not. The item response theory (IRT) models used to summarize performance in a subject area or subarea posit an unobservable proficiency variable θ to summarize performance on the items in that area. The fact that θ values are not observed even for the respondents in the sample requires additional statistical machinery to draw inferences about θ distributions and to quantify the uncertainty associated with those inferences. To this end, we have adapted Rubin's (1987) "multiple imputations" procedures for missing data to the context of latent variable models to produce the "plausible values" that appear in the NAEP 1990 secondary-use data files.

The essential idea of plausible values methodology is that even though we do not observe the θ value of respondent i, we do observe other kinds of variables that are related to it: x_i, the respondent's answers to the cognitive items he or she was administered in the area of interest, and y_i, the respondent's answers to demographic and background variables. Suppose we would like to draw inferences about a number $T(\Theta,Y)$ that could be calculated explicitly if the θ and y values of each member of the population were known. Suppose further that we would be able to estimate T from a sample of N pairs of θ and y values by the statistic $t(\theta,y)$, where $(\theta,y) = (\theta_1,y_1,\ldots,\theta_N,y_N)$, and that we could estimate the variance in t around T due to sampling respondents by the function $U(\theta,y)$. Given that observations consist of (x_i,y_i) rather than (θ_i,y_i), we can approximate t by its expected value conditional on (x,y), or

$$\begin{aligned} t^*(x,y) &= E[t(\theta,y)\,|\,x,y] \\ &= \int t(\theta,y)p(\theta\,|\,x,y)\,d\theta \ . \end{aligned}$$

It is possible to approximate t^* with random draws from the conditional distributions $p(\theta_i\,|\,x_i,y_i)$, which are obtained for all respondents by the method described in Chapter 7. Let $\hat{\theta}_m$ be the m^{th} such vector of "plausible values." It is a plausible representation of what the true θ might have been, had we been able to observe it. The following steps describe how an estimate of a scalar statistic $t(\theta,y)$ and its sampling variance can be obtained from M (>1) such sets of plausible values. (Note: five sets are provided on the data files for each subject area or subarea analyzed by these procedures.)

1) Using each set of plausible values $\hat{\theta}_m$ in turn, evaluate t as if the plausible values were true values of θ. Denote the results $\hat{t}_m$, for m=1,...,M.

2) Using the multiple weight jackknife approach, compute the estimated sampling variance of $\hat{t}_m$, denoting the result as U_m.

3) The final estimate of t is

$$t^* = \sum_{m=1}^{M} \hat{t}_m \,/\, M \ .$$

4) Compute the average sampling variance over the M sets of plausible values, to approximate uncertainty due to sampling respondents:

$$U^* = \sum_{m=1}^{M} U_m \,/\, M \ .$$

5) Compute the variance among the M estimates $\hat{t}_m$, to approximate uncertainty due to not observing θ values from respondents:

$$B_M = \sum_{m=1}^{M} (\hat{t}_m - t^*)^2 \,/\, (M-1) \ .$$

6) The final estimate of the variance of t^* is the sum of two components:

$$V = U^* + (1 + M^{-1})\,B_M \ .$$

(Note: NAEP reports use a single jackknife estimate U_m in place of the average of five, as would be required for U^*; see section 8.5.)

Suppose that the statistic $[t(\theta,y) - T]/U^{1/2}$ would follow a t-distribution with d degrees of freedom. Then the distribution of $(t^* - T)/V^{1/2}$ is also approximately t, with degrees of freedom given by

$$\nu = (M - 1)(1 + r_M^{-1}) \frac{d}{d + r_M^{-2}(M-1)}$$

where r_M is the relative increase in variance due to not observing θ values:

$$r_M = (1 + M^{-1})\, B_M \,/\, U^* \ .$$

When B is small relative to U, and d is large, a normal approximation suffices. This is the case with main NAEP reporting variables, and the normal approximation is routinely applied to flag "significant" results.

For k-dimensional t, such as the k coefficients in a multiple regression analysis, each U_m and U^* are covariance matrices, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity

$$(T - t^*)\, V^{-1}\, (T - t^*)'$$

is approximately F distributed, with degrees of freedom equal to k and ν, with ν defined as above but with a matrix generalization of r_M:

$$r_M = (1 + M^{-1})\, Trace(B_M U^{*-1}) \,/\, k \ .$$

By the same reasoning as used for the normal approximation for scalar t, a chi-square distribution on k degrees of freedom often suffices.

Computation of statistics t^* involving the proficiency of a single subject area, or the composite value in mathematics or science, and categories of variables included in the conditioning variables y (described in Chapter 7), yields consistent estimates of the corresponding population values T. *Statistics involving background variables y that were not conditioned on, or relationships among proficiencies from different subject areas, are subject to biases whose magnitudes depend on the type of statistic and the strength of the relationships of the nonconditioned background variable(s) to the variables that were conditioned on and to the proficiency of interest. The direction of the bias is typically to underestimate the effect of nonconditioned variables.*[1] For a given statistic t^* involving one subject area and one or more nonconditioned background variables, the magnitude of the bias is related to the fraction of information about T that is missing because θ is not observed:

---

[1] For details, see section 10.3.5 of *Implementing the New Design: The NAEP 1983-84 Technical Report* and section 8.4.3 of *Expanding the New Design: The NAEP 1985-86 Technical Report.*

$$\gamma_M = \frac{r_M + 2/(\nu + 3)}{r_M + 1} \quad .$$

8.5 APPROXIMATIONS

The major computational load in calculating uncertainty measures for any statistic exists in the computation of the uncertainty due to sampling variability. As noted in the last section, a jackknife estimate of the variability of a statistic based on one or more observed NAEP variables in the 1990 Trial State and national assessment samples requires computing the statistic 57 times. This section describes a less computationally intensive approximation to sampling variability of any statistic.

As indicated in section 8.3.1, it is inappropriate to estimate the sampling variability of any statistic based on the NAEP database by using simple random sampling (SRS) formulas. These formulas, which are the ones used by most standard statistical software such as SPSS and SAS, will produce variance estimates that are generally much smaller than is warranted by the sample design.

It may be possible to account approximately for the effects of the sample design by using an inflation factor, the design effect, developed by Kish (1965) and extended by Kish and Frankel (1974). The design effect for a statistic is the ratio of the actual variance of the statistic (taking the sample design into account) over the simple random sampling variance estimate based on the same number of elements. The design effect may be used to adjust error estimates based on simple random sampling assumptions to account approximately for the effect of the design. In practice, this is often accomplished by dividing the total sample size by the design effect and using this effective sample size in the computation of errors. Note that the value of the design effect depends on the type of statistic computed and the variables considered in a particular analysis as well as the clustering effects occurring among sampled elements and the effects of any variable weights resulting from variable overall sampling fractions.

On the basis of empirical results and theoretic considerations, Kish and Frankel (1974) have developed several conjectures about design effects:

1) Generally, the design effects for complex statistics from complex samples are greater than 1, causing variances based on simple random sampling assumptions to tend to be underestimated.

2) The design effects for regression coefficients tend to be smaller than the corresponding design effects for means of the same variables. Hence, these latter estimates, which are more easily computed, tend to overestimate the design effects of complex statistics. For correlation coefficients and partial correlation coefficients, the design effect for the mean should be used (Skinner, Holt, & Smith, 1989, p. 70).

3) The size of the design effects of complex statistics tends to parallel those of means; variables with a high design effect of the mean also tend to have high design effects for complex statistics involving those variables.

To incorporate the design effect idea in a statistical analysis, proceed in the following manner:

1) For a given class of statistics (e.g., means, proportions, regression coefficients), compute the jackknife variance described in section 8.3.1 for a number of cases. The cases should cover the range of situations for which the approximation is to be used. If various subpopulations are to be considered, it is important to have information on the relative variability within each subgroup. This is especially important if certain subgroups are more highly clustered in the sample.

2) For the identical cases, compute the simple random sampling variance given the elements in the sample. To account properly for the difference between the number of individuals being sampled and the total of the sampling weights, the weights should be scaled so that their sum equals the sample size.

3) For each case, compute the design effect where the design effect for case j is

$$deff_j = Var_{JK}(t_j) / Var_{CON}(t_j) \quad ,$$

the ratio of the jackknife variance estimate of the statistic to its simple random sampling variance estimate.

4) If the design effects for the various cases are tolerably similar, choose an overall composite design effect. If the design effects for certain subgroups appear to cluster around a markedly different value from the remaining cases, treat those subgroups separately.

5) In the case that a consistent overall design effect has been found:

   a) Rescale the weight of each individual so that the sum of the scaled weights is equal to the effective sample size

   $$N_{eff} = \frac{\text{sample size}}{\text{design effect}}$$

   (that is, multiply each weight by N_{eff}/W_{TOT}, where W_{TOT} is the sum of the original weights).

   b) Conduct a traditional weighted analysis using these scaled weights.

6) The degrees of freedom for any variance estimates obtained by using this approach is still at best 56, the number of pairs, as it was for the jackknife. Accordingly, tests of

significance produced by standard programs (which will use the effective sample size minus the number of parameters for error degrees of freedom) should be interpreted with extreme caution because they are likely to be too liberal. Significance and inferential procedures will be improved if based on the smaller error degrees of freedom that were indicated for the degrees of freedom for the jackknife variance estimate (section 8.3.1), although this estimate of effective degrees of freedom will also be an overstatement, since the paired variables will not be normally distributed.

8.6 A NOTE CONCERNING MULTIPLE COMPARISONS

If many statistical tests are conducted at one time, it is likely that significance tests will overstate the degree of statistical significance of the results. In the preceding sections, we noted that because of the design of the NAEP sample, conventional significance tests will overstate significance, because they fail to consider the effects of clustering. In contrast, the problem of multiple comparisons noted here is independent of sample design; it arises even if one use the appropriate statistical tests described previously. The problem arises because the more statistical tests are calculated, the more likely it becomes that one will find a "significant" finding because of chance variation. In other words, the chance of a type I error -- a spurious "significant" finding -- rises with the number of tests conducted.

More technically, if J multiple hypothesis tests are performed, each with a type I error rate (the probability of rejecting the null hypothesis when the null hypothesis is true) of α, the type I error rate for the entire set of tests could be as high as $J\alpha$. Therefore, it is desirable to use a multiple comparison procedure to control the overall error rate for the entire set of hypothesis tests. In the present case, it is advantageous to use a procedure that allows control of the error rate for sets of varying size that may include both pairwise and complex comparisons. (An example of a complex contrast is a comparison of one group to the average of two other groups.) The Dunn-Bonferroni approach is, therefore, a good choice. To apply this method in its simplest form, we need only decide at what level we wish to control the setwise error rate (α_s) and then set the type I error rate for each comparison equal to $\alpha_C = \alpha_s/J$, where J is the number of comparisons.

For example, suppose we wanted to perform three pairwise comparisons between regional groups, as well as one complex comparison, controlling α_s at .05. The type I error rate for each comparison should be set at $\alpha_C = \alpha_s/J = .05/4 = .0125$. The required critical value can be obtained from a table of the Bonferroni t-statistic (Miller, 1981, p. 238) with the appropriate degrees of freedom.

Chapter 9

CONTENT AND FORMAT OF DATA FILES, LAYOUTS, AND CODEBOOKS

Chapter 9: CONTENT AND FORMAT OF DATA FILES, LAYOUTS, AND CODEBOOKS

9.1 INTRODUCTION

This chapter describes in detail the content and format of each 1990 data file and the accompanying printed layouts and codebooks.

Each data package contains a data file for each student sample and questionnaire instrument. Three other types of files are provided for each data file: a set of SPSS-X control statements for generating an SPSS-X system file; a set of SAS control statements for generating a SAS system file; and a machine-readable catalog file containing parameter data and information for each field in a data record.

The accompanying printed documentation contains a file layout and data codebook for each data file. Each layout contains the essential processing and labeling information on one line for each data field. Each codebook contains more descriptive information for each field.

9.2 DATA FILES

There are four file types for each sample administered in the 1990 assessment. The files are arranged by sample within file type. The file order, names, and characteristics are given in Table 9-1. The files are named according to the following convention:

The first index level (up to the first period) designates the sample:

TSASTUD	State student sample
TSASCHL	State school sample
TSAEXCL	State excluded student sample
NWPSTUD	National winter public-school student sample
NWPSCHL	National winter public-school school sample

The second index level is the file type:

DAT	The raw data file
SAS	The SAS control statements for generating a SAS system file
SPX	The SPSS-X control statements for generating an SPSS-X system file
CAT	A machine-readable catalog of item and variable information

Table 9-1

NAEP 1990 State Data Package Description

Files	Record Length	Number of Records	File Name
Data Files			
1. State Student Sample	1701	*	TSASTUD.DAT
2. State School Sample	755	*	TSASCHL.DAT
3. State Excluded Student Sample	561	*	TSAEXCL.DAT
4. National Winter Public-school Student Sample	1365	2879	NWPSTUD.DAT
5. National Winter Public-school School Sample	814	95	NWPSCHL.DAT
SPSS-X Control Statement Files			
6. State Student Sample	80	2138	TSASTUD.SPX
7. State School Sample	80	488	TSASCHL.SPX
8. State Excluded Student Sample	80	366	TSAEXCL.SPX
9. National Winter Public-school Student Sample	80	2143	NWPSTUD.SPX
10. National Winter Public-school School Sample	80	558	NWPSCHL.SPX
SAS Control Statement Files			
11. State Student Sample	80	1067	TSASTUD.SAS
12. State School Sample	80	371	TSASCHL.SAS
13. State Excluded Student Sample	80	254	TSAEXCL.SAS
14. National Winter Public-school Student Sample	80	1023	NWPSTUD.SAS
15. National Winter Public-school School Sample	80	400	NWPSCHL.SAS
16. Format Library Generator	80	615	FORMATS.SAS
Machine-readable Catalog Files			
17. State Student Sample	1328	608	TSASTUD.CAT
18. State School Sample	1328	248	TSASCHL.CAT
19. State Excluded Student Sample	1328	156	TSAEXCL.CAT
20. National Winter Public-school Student Sample	1328	574	NWPSTUD.CAT
21. National Winter Public-school School Sample	1328	269	NWPSCHL.CAT

---

* Number of records varies by state.

9.2.1 Raw Data Files

Depending on the sample, each raw data file contains one record per student, excluded student, or school. All raw data files are rectangular, that is, record lengths are fixed and a given variable always occurs in the same position on every record within a file.

The NAEP data files are structured to facilitate matching among the three samples (student, excluded student and school). The teacher data has already been linked with the appropriate students on the state and national samples. For the purposes of analysis and reporting, only two types of linkages are valid:

1) school with student and teacher (state and national)
2) school with excluded student (state only)

The primary linkage on the state files is through the school code fields: SCH, SSCH, and XSCH. The linkage between the national files is accomplished through the primary sampling unit/school code fields: PSUSCH and SPSUSCH. All files are ordered by these fields to permit direct match-merging without the need to reorganize.

Because of the nature of the BIB spiral design (see Chapter 3), students were administered different blocks of items during the 1990 assessment. As a result, each student record contains blank spaces for the item blocks that were not included in the student's assessment booklet (missing by design). Fields are also blank for items that did not appear in booklets because of a printing error (e.g., incorrect block in booklet, missing pages) and for the professionally scored items that were not included in reliability checks (see section 5.4 in Chapter 5). Additionally, items that were either missed by the scorers or given erroneous codes by the scorers were coded as blank fields.

Special codes (Table 9-2) were assigned to item response fields for which "I don't know," omitted, not reached, or multiple responses were indicated.

Table 9-2
Special Response Codes

Code (Width=1)	Code (Width=2)	Definition
7	77	I DON'T KNOW (multiple-choice items)
8	88	OMITTED
9	99	NOT REACHED
0	0	MULTIPLE RESPONSE (multiple-choice items)

9.2.2 SPSS-X and SAS Control Statement Files

All data files in the data package are accompanied by separate control files to facilitate the creation of SPSS-X and SAS system files. These control files include statements for variable definitions, variable labels, missing value codes, value labels, and an optional section for creating and storing scored variables. Each set of control statements also generates unweighted descriptive statistics of the reporting variables for the related data file and a listing of the contents of the saved system file.

Specific details on the structure and use of these control files are provided in Chapter 10.

9.2.3 Machine-readable Catalog Files

The machine-readable catalog files are designed primarily for users who want to use a programming language or package other than SAS or SPSS-X to analyze the data. These files may also be processed by SAS or SPSS-X to produce listings or informational reports.

Each catalog file contains a record for each variable or item on its corresponding data file. Table 9-3 contains the machine-readable catalog data layout. Specific information concerning the contents of the catalogs is provided below.

FIELD SEQUENCE NUMBER	Fields are numbered sequentially to represent the order in which they appear on the raw data record.
FIELD NAME	A short name of up to eight characters that uniquely identifies the field.
START COLUMN	The start location of the field on the data record.
END COLUMN	The end location of the field on the data record.
FIELD WIDTH	The number of column positions used by the field.
DECIMAL PLACES	The number of digits to the right of the decimal point in the field. The raw data contain implicit decimal points.
FIELD TYPE	The files include two field types: Type 1 (discrete) fields designate discrete data with a fixed number of responses. Type 1 fields may include raw item responses or imputed categorical variables. Type 2 (continuous) fields designate continuous numerical data without fixed ranges.
MINIMUM VALID RESPONSE	The minimum value of valid responses for an item, excluding "I don't know" responses.

Table 9-3
NAEP 1990 State Machine-Readable Catalog File Layout

Start and End Columns	Field Width	Field Type	Field Description	Comments
1 - 4	4	N	Field Sequence Number	New NAEP Ident.
5 - 12	8	A	Field Name	
13 - 16	4	N	Start Column	
17 - 20	4	N	End Column	
21 - 22	2	N	Field Width	
23 - 23	1	N	Decimal Places	
24 - 24	1	N	Field Type	
25 - 27	3	N	Minimum Valid Response	
28 - 30	3	N	Maximum Valid Response	
31 - 32	2	N	Minimum Correct Response	
33 - 34	2	N	Maximum Correct Response	
35 - 36	2	N	I Don't Know (IDK) Code	
37 - 38	2	N	Omit Code	
39 - 40	2	N	Not Reached Code	
41 - 42	2	N	Multiple Response Code	
43 - 92	50	A	Field Label	
93 - 108	16	A	Old NAEP Identification	
109 - 109	1	N	Calculator Category	
110 - 114	5	N	Scaling Categories	
115 - 138	24	N	IRT Parameters	
139 - 140	2	N	Number of Data Codes and Labels	
141 - 142	2	N	Code Value	1st Data Code
143 - 162	20	A	Code Label	
163 - 164	2	N	Code Value	2nd Data Code
165 - 184	20	A	Code Label	
.	.	.	.	
.	.	.	.	
1307 - 1308	2	N	Code Value	24th Data Code
1309 - 1328	20	A	Code Label	

MAXIMUM VALID RESPONSE	The maximum value of valid responses for an item, excluding "I don't know" responses.
MINIMUM CORRECT RESPONSE	For scorable cognitive items, the minimum or only correct response value.
MAXIMUM CORRECT RESPONSE	For scorable cognitive items with more than one correct response, the maximum correct response value. For example, if possible responses for a professionally scored item ranged from 0 to 5, and 3 to 5 were considered acceptable responses, the first position of the field would contain a 3 and the second position would contain a 5.
I DON'T KNOW RESPONSE CODE	For multiple-choice items, the value in the "IDK Value" column is the code assigned to "I don't know" responses when that option was given.
OMIT CODE	The value in this field is the code assigned to nonresponses for the following types of items: 1) All noncognitive items (background, attitude, and questionnaire) 2) Cognitive items that are followed by valid responses to other items in the same block
NOT REACHED CODE	The value in this field is the code assigned to nonresponses to cognitive items after the last valid response in a block.
MULTIPLE RESPONSE CODE	The value in this field is the code assigned to multiple-choice items for which the subject indicated more than one response.
FIELD LABEL	A 50-character description of the item or variable.
OLD NAEP ID	The identification code previously assigned to the item if the item had been used before 1983.
CALCULATOR CATEGORY	For mathematics cognitive items, denotes appropriateness of calculator usage: 1 = Calculator-active 2 = Calculator-inactive 3 = Calculator-neutral
SCALING CATEGORIES	Each column corresponds to one of the five major scales derived for the 1990 assessment and contains either a zero (0) or one (1). Each nonzero entry in a column indicates usage of the item on a scale. Table 9-4 lists these column locations, their corresponding scales, and the name of the scale variable(s).

IRT PARAMETERS	Three eight-character fields containing the IRT item parameters: "a" (discrimination); "b" (difficulty); and "c" (asymptote). Each parameter is represented to a precision of five decimal places with an explicit decimal point.
NO. OF DATA CODES AND LABELS	The number of valid data codes. For item responses, these include illegible, off-task, "I don't know", nonrateable, omit, not reached, and multiple responses.
DATA CODES AND LABELS	For each valid discrete variable, a two-position field that shows the data code and a 20-position text field that provides a brief description of the code. There can be up to 54 codes; if there are fewer than 54, the remaining fields are blank.

Table 9-4
Scaling Categories and Codes

Column	Scale	Subscale	Scaling Variables
110	Mathematics	Numbers and Operations	MRPSCAx
111	Mathematics	Measurement	MRPSCBx
112	Mathematics	Geometry	MRPSCCx
113	Mathematics	Data Analysis, Statistics, and Probability	MRPSCDx
114	Mathematics	Algebra and Functions	MRPSCEx

9.3 PRINTED DOCUMENTATION

Each state's data files are accompanied by a book containing the layouts and codebooks for each data file. These documents are grouped by layout and codebook pair in the same order as the data files.

9.3.1 File Layouts

Each file layout includes the following information for each data field:

SEQ. NO.	Sequence number. Fields are numbered sequentially to represent the order in which they appear on the data record.

FIELD NAME	A short name (of up to eight characters) that identifies the field. This name is used consistently across all documentation, SAS & SPSS-X control files, and catalog files to identify each field uniquely within a data file. In general, nonresponse data field names are abbreviations of the field descriptions. Field names associated with response data are formatted as follows:

Position 1	identifies nature/source of the response data: B = Common background item within common background block S = Subject-related background or attitude item (noncognitive mathematics items from the 1984 and 1986 assessments) N = Cognitive item within cognitive block (cognitive mathematics items from the 1984 and 1986 assessments) M = Mathematics cognitive or background item T = Teacher questionnaire item C = School questionnaire item X = Excluded student questionnaire item
Positions 2 - 5	identify an exercise (student files) or question (school, teacher, excluded student files). Mathematics background items are identified by "M" in position 1 and "8" in position 2.
Positions 6 & 7	identify a part within an exercise (student files) or a part within a question (school, teacher, excluded student files).
Position 8	identifies the block containing an item (Student files only). The numeric designation (2 through 9) has been replaced by an alphabetic one (B through I). This position is blank for questionnaire items and all other variables.

COL. POS.	Column position. The start location of the field on the data record.
FIELD WIDTH	The number of column positions used by the field.

DECIMAL PLACES	The number of digits to the right of the decimal point in the field. The raw data contain implicit decimal points.
TYPE	The files include five field types: Type C (continuous) fields designate continuous numerical data without fixed ranges. Type D (discrete) fields designate discrete data with a fixed number of responses. Type D fields may include raw item responses or imputed (derived) categorical variables. Type DI (discrete with "I don't know") fields designate discrete data with a special code for "I don't know" responses. Type O (open-ended) fields designate free-response items in the student data that were professionally scored at ETS.
RANGE	The range of values or of valid responses for a field.
KEY VALUE	The correct response for a multiple-choice cognitive item. For those open-ended items that were scored using a cut-point scale, the key is expressed as a range of values.
SHORT LABEL	A brief description of the information in the field.

9.3.2 Codebooks

The codebook contains one or more lines of information for each data field, depending on the data type. The first line of each codebook entry contains the following information:

SEQ. NO.	Sequence number. In conjunction with the numbers assigned in the layouts, the fields in the codebooks are numbered sequentially.
FIELD NAME	A short name of up to eight characters that uniquely identifies the field. If an item was used in an assessment prior to 1984, its old identification number is printed in parentheses just below the FIELD NAME.
REL. IND.	Release indicator. Indicates that an item is available for unrestricted public use (P) or is non-released and reserved exclusively for use by NAEP (N).
TYPE	In conjunction with the five field types defined for the layouts, the field type is designated as continuous (C), discrete (D), discrete with "I don't know" (DI), or open-ended (O).

BLOCK	For assessment items, indicates the block in which an item appeared for the cohort of students for which the codebook was prepared.
ITEM NO.	Indicates the order of an item within a block for the grade/age group of students for which the codebook was prepared.
AGES	Indicates the student grade/age or age groups to which an item was administered, as follows: Main Sample 1 = Grade 4/Age 9 2 = Grade 8/Age 13 3 = Grade 12/Age 17
NAME/ DESCRIPTION	A brief description of the information in the field. *NOTE: To maintain item security, the text for the NAME/DESCRIPTION field and responses (data value labels) has been replaced by short descriptions for items classified as non-released.*

For all discrete and open-ended data fields, the third and subsequent lines contain each valid data value, its associated label, and the unweighted frequency of that value in the data file. (For cognitive items, the correct response code is indicated by an asterisk.) The last line under each discrete variable entry contains the "TOTAL" or sum of the frequency counts as an extra check for analyses.

If an item was used in IRT scaling, its scale identification and parameter values are listed to the right of the frequency data. The first column contains the code for the scale for which the item was calibrated, the second column denotes the IRT parameter type, and the third column contains the parameter values. The scale codes and their meaning are as follows:

Code	Scale
N&O	Numbers and Operations
MEA	Measurement
GEO	Geometry
DAS	Data Analysis, Statistics, and Probability
ALG	Algebra and Functions

Chapter 10

WORKING WITH SPSS-X AND SAS

Chapter 10: WORKING WITH SPSS-X AND SAS

10.1 INTRODUCTION

This chapter discusses the use of the statistical software SPSS-X and SAS in analyzing 1990 NAEP data. Included are procedures for creating SPSS-X and SAS system files, merging files using SPSS-X and SAS, using the jackknife procedure with SPSS-X and SAS to estimate standard errors, and an example using NAEP data with SAS.

10.2 SPSS-X AND SAS CONTROL STATEMENT FILES

All data files in the NAEP data package are accompanied by separate control files to facilitate the creation of SPSS-X and SAS system files. These control files include statements for variable definitions, variable labels, missing value codes, value labels, and an optional section for creating and storing scored variables. Each set of control statements also generates unweighted descriptive statistics of the reporting variables for the related data file and a listing of the contents of the saved system file.

Users who are performing analyses using data contained on magnetic tape should be aware that the system file generation programs cannot run if both the control statement file and its corresponding data file reside on the same tape. Both SPSS-X and SAS will try to read a data file before they have completed processing the control statement file, which is physically impossible if both files are on the same tape. The user is advised to copy the control files to disk, as they require less storage space and allow the user to edit the control statements before generating the system files.

The common features of both types of control files, as well as general guidelines, are provided below.

VARIABLE DEFINITION	The field names are listed in the order in which they appear on the file, along with their column position and input formats. If the field is numeric with no decimal places, no format is provided. Otherwise, the format is indicated by a number for the number of decimal places, or by '$' or '(A)' for a nonnumeric field.
VARIABLE LABELS	A 40-character text description for each field.
MISSING VALUES	All blank fields in the data are automatically set to the system missing value by each package. However, all multiple-choice and some open-ended items were prone to either multiple or out-of-range responses. These items were coded as fields of nines in the data files. The control statement files instruct each system to treat these values as missing.

VALUE LABELS	All numeric fields with discrete (or categorical) values are provided with 20-character text descriptors for each value within the variable's range. The value labels, or formats, for the SAS control statements have been pooled across all three samples into a file for one-time processing and loading into a SAS format library. A listing that links the field names to the SAS format names is provided in each codebook.
SCORING	For each item with one or more correct responses, control statements are provided for creating a scored variable, its label, and its value labels. The scoring of each item is performed according to the following scheme: missing values are copied as is; correct response values are recoded to 1; all other values, including no response and "I don't know," are recoded as 0. The scoring of the omit, not reached, and "I don't know" values are coded separately from other incorrect responses to allow the user to edit these control statements and substitute alternate values. Each scorable item is replaced by its scored value, along with its new value labels and missing value declarations. The entire block of scoring control statements is performed conditionally by default and will not be saved on the output system file. To save the scored variables permanently, the user must edit the control statement file and make changes to a few specified statements. It is not possible under this scheme to save both the raw and scored responses to the same item.

10.3 CREATING SPSS-X SYSTEM FILES

Each SPSS-X control statement file is linked to its corresponding data file through the file name: the suffix DAT in the data file name is replaced by SPX to obtain the control statement file name. For example, file TSASTUD.SPX is the control statement file for data file TSASTUD.DAT.

All SPSS-X control statement files have been generated according to the structure in Table 10-1.

The TEMPORARY command instructs SPSS-X to perform the subsequent scoring statements on a temporary basis and delete the new variables after the next procedure encountered (FREQUENCIES). Thus, the scored variables will NOT be saved on the system file unless the TEMPORARY command is commented or edited out.

All control statement files assume that the file handle (or DDNAME) for the input data file is RAWDATA, and the file handle for the output system file is SYSFILE.

The control statements were coded according to the command and procedure descriptions in the *SPSS Reference Guide* (SPSS, Inc., 1990). They were tested under SPSS-X Version 4.1 (IBM-OS/MVS).

Table 10-1

SPSS-X Control Statement Synopsis

```
TITLE
      label for sysout of file generation run
FILE LABEL
      label to be stored with file
DOCUMENT
      text description of data to be saved in file
DATA LIST FILE=RAWDATA
      variable names, locations, and formats
VARIABLE LABELS
      40-character label for each variable
MISSING VALUES
      list of variables to have user-missing values assigned
VALUE LABELS
      variable names, values, and value labels
TEMPORARY         ** delete this statement to save scored variables **
RECODE
      oldvar (SYSMIS=SYSMIS)  (0=9)  (keyval=1)
      (nrval=0)  (omval=0) (idkval=0)  (ELSE=0)
.
.
.

MISSING VALUES
      for recodes of multiple responses
VALUE LABELS
      1=Correct    0=Incorrect
FREQUENCIES
      reporting variables
SAVE   OUTFILE=SYSFILE/COMPRESSED
DISPLAY LABELS
```

BEST COPY AVAILABLE

10.4 CREATING SAS SYSTEM FILES

Each SAS control statement file is linked to its corresponding data file through the file name: the suffix DAT in the data file name is replaced by SAS to obtain the control statement file name. For example, file TSASTUD.SAS is the control statement file for data file TSASTUD.DAT.

All SAS control statement files have been generated according to the structure in Table 10-2. They use the SAS Macro Language facility to reduce the number of source statements generated and provide consistent performance of repetitive functions. Therefore, the user must ensure that the MACRO option is invoked when processing any of the control statement files.

The DO OVER through END statements following each ARRAY statement set up the conversion of the "I Don't Know," omit, not reached, and multiple response codes to the system missing value. However, once this conversion is executed and saved on the system file, these recoded values will be indistinguishable from actual missing values on the original data file. For this reason, these statements have been commented out to allow the user to decide which, if any, of the values are to be recoded. To activate the recoding, delete the asterisks preceding the DO OVER and END statements and from the appropriate IF THEN statement(s).

The missing value transformations are followed by a series of SAS macro definitions for scoring the cognitive items. The RECODE macro is used by the SCORE macro to transform the responses to each item into score values. The RECODE macro may be edited by the user to transform the special codes for each item consistently into other values.

At the end of the control statements, the SCORE macro is commented out. To save the scored variables on the system file, the user should uncomment the %SCORE statement.

A separate file of SAS control statements is provided that contains the SAS formats to be used by all variables. This file, named FORMATS.SAS, may be executed before all other SAS control statement files, and does not require a raw data file for input. The format specifications will be saved in a library designated to the system as SASLIB. Each codebook contains a list of all discrete variables and the format values to be used in any SAS analysis.

The control statements were coded according to the command and procedure descriptions in the *SAS Language: Reference, Version 6, First Edition* (SAS Institute, Inc., 1990). They were tested under SAS Release 6.06 (IBM-OS/MVS).

10.5 MERGING FILES UNDER SPSS-X OR SAS

The NAEP data files are structured to facilitate matching among the four instruments (student, excluded student, teacher, and school). The teacher questionnaire has already been linked with the appropriate students from the state and national samples. For the purposes of analysis and reporting, only two types of linkages are valid:

1) school with student and teacher
2) school with excluded student

Table 10-2

SAS Control Statement Synopsis

```
TITLE
DATA SYSFILE.xxx;
INFILE RAWDATA;
INPUT
      variable names, positions, and formats
LABEL
      40-character variable labels

*ARRAY DKn (I)                list of variables with "I Don't Know"
*DO OVER DKn;                 codes to be recoded for missing
*  IF DKn=7   THEN DKn=.;
*  END;
*ARRAY OMn (I)                list of variables with omit codes to be
*DO OVER OMn;                 recoded for missing
*  IF OMn=8   THEN OMn=.;
*  END;
*ARRAY NRn (I)                list of variables with not-reached codes
*DO OVER NRn;                 to be recoded for missing
*  IF NRn=8   THEN NRn=.;
*  END;
*ARRAY MRn (I)                list of variables with multiple response
*DO OVER MRn;                 codes to be recoded for missing
*  IF MRn=9   THEN MRn=.;
*  END;
LENGTH DEFAULT=2
      other variables with appropriate lengths;
%MACRO RECODE;
      SAS macro to perform scoring for each variable
%MEND RECODE;
%MACRO SCORE;
%RECODE     (oldvar,newvar,idkval,nrval,keyval)
      .
      .
%MEND SCORE;
*%SCORE                 ** delete asterisk to save scored variables **
PROC FREQ;
TABLES
      reporting variables
PROC CONTENTS NOSOURCE POSITION;
```

The primary linkage on all files is through the school code fields: SCH, SSCH, and XSCH. All files are organized by these fields to permit direct match-merging without the need to re-sort.

When a hierarchical file match is performed, both SPSS-X and SAS build a rectangular file at the level of the lowest file in the match. Each record from the higher order file is repeated across the corresponding records of the lower order file. For example, in matching school with student data, the information from one school record is repeated across all student records belonging to that school. Clearly, the number of variables from the higher order file will have a greater impact on the size of the resulting merged file.

The examples shown in Tables 10-3 and 10-4 will perform direct matches according to the two linkages listed above. The KEEP statements are not necessary to the performance of the merge, but when they are applied to only those variables required for analysis, they will make more efficient use of computer resources. These examples also assume that no transformations are to be performed on the input files. If transformations are desired for analysis, the most efficient course to follow would be to transform the variables from the higher order file first, perform the match procedure, then transform the variables from the lower order file.

10.6 COMPUTING THE ESTIMATED VARIANCE OF A MEAN (JACKKNIFING) USING SPSS-X OR SAS

This section presents the two multiweight methods for computing the estimated variance of a mean in SPSS-X and SAS program code form (see section 8.3 in Chapter 8 for a discussion of the jackknife procedure). The first method may be used for any variable except the plausible values for mathematics. The second method, which should be used for the plausible values, employs a correction for the variance in estimating the values (correction for imputation).

For each variable to be jackknifed, generate two vectors of weighted sums and products. Sum these vectors across the entire file using the AGGREGATE (SPSS-X) or SUMMARY (SAS) procedures. From the weighted sums compute the weighted means and then compute the estimated variance and standard error.

One advantage to this approach is that it will accomplish the computation in one pass of the data. Another advantage, afforded by the AGGREGATE (SPSS-X) and SUMMARY (SAS) procedures, is the facility to compute subgroup statistics by using the BREAK keyword (SPSS-X) or CLASS option (SAS) with the variable(s) defining the subgroups. All computations performed subsequent to the aggregation procedure are performed on each record of the collapsed file, corresponding to one of the subgroups. In the examples in Tables 10-5, 10-6, 10-7, and 10-8, the variable DSEX is used as a break control variable, and the derived statistics are printed for each gender code.

Table 10-3

Matching School and Student Files

SPSS-X

```
MATCH FILES
      TABLE=SCHOOL/
            RENAME=(SSCH=SCH)/
      FILE=STUDENT/
            KEEP=SCH,other school & student variables/
      BY=SCH.
```

SAS

```
DATA MATCH1;
      MERGE SCHOOL(RENAME=(SSCH=SCH)
                  KEEP=SSCH other school variables)
            STUDENT(KEEP=SCH other student variables);
      BY SCH;
```

Table 10-4

Matching School and Excluded Student Files

SPSS-X

```
MATCH FILES
      TABLE=SCHOOL/
            RENAME=(SSCH=SCH)/
      FILE=EXCLUDE/
            RENAME=(XSCH=SCH)/
            KEEP=SCH,other school and excluded student variables/
      BY=SCH.
```

SAS

```
DATA MATCH3;
      MERGE SCHOOL (RENAME=(SSCH=SCH)
                  KEEP=SSCH other school variables)
            EXCLUDE(RENAME=(XSCH=SCH)
                  KEEP=SCH other excluded student variables);
      BY SCH;
```

Table 10-5

Standard Error Computation: Multiweight Method Using SPSS-X

```
GET FILE=SYSFILE/                   (System file for sample)
      KEEP=DSEX,WEIGHT,SRWT01 TO SRWT56,X.
VECTOR WT=SRWT01 TO SRWT56.
VECTOR WX(56).
SELECT IF (NOT SYSMIS(X)).
COMPUTE WTX=WEIGHT*X.
LOOP #I=1 TO 56.
   COMPUTE WX(#I) = WT(#I)*X.
END LOOP.
AGGREGATE  OUTFILE=*/BREAK=DSEX/UWN=N(WEIGHT)/
      SWT,SW1 TO SW56 = SUM(WEIGHT,SRWT01 TO SRWT56)/
      SWX,SX1 TO SX56 = SUM(WTX,WX1 TO WX56)/.
VECTOR SW = SW1 TO SW56.
VECTOR SX = SX1 TO SX56.
COMPUTE XBAR = SWX/SWT.
COMPUTE XVAR = 0.
LOOP #I=1 TO 56.
   COMPUTE #DIFF = SX(#I)/SW(#I) - XBAR.
   COMPUTE XVAR = XVAR + #DIFF * #DIFF.
END LOOP.
COMPUTE XSE = SQRT(XVAR).
PRINT FORMATS XVAR,XSE (F8.4).
LIST VARIABLES=DSEX,UWN,SWT,XBAR,XVAR,XSE.
FINISH.
```

Table 10-6

Standard Error Computation: Multiweight Method Using SAS

```
DATA A;
      SET SYSFILE.TSASTUD;
      ARRAY WT SRWT01-SRWT56;
      ARRAY WX WX1-WX56;
      IF (X NE .);
      WTX = WEIGHT*X;
      DO OVER WT;
            WX = WT*X;
      END;
PROC SUMMARY;
      CLASS DSEX;
      VAR WEIGHT SRWT01-SRWT56 WTX WX1-WX56;
      OUTPUT OUT=B    N(WEIGHT)=UWN
            SUM(WEIGHT WTX SRWT01-SRWT56 WX1-WX56)=
                SWT SWX SW1-SW56 SX1-SX56;
DATA C;
      SET B;
      ARRAY SW SW1-SW56;
      ARRAY SX SX1-SX56;
      XBAR = SWX/SWT;
      XVAR = 0;
      DO OVER SW;
            DIFF = (SX/SW)-XBAR;
            XVAR = XVAR+DIFF*DIFF;
      END;
      XSE = SQRT(XVAR);
PROC PRINT;
      VAR DSEX UWN SWT XBAR XVAR XSE;
```

Table 10-7

Standard Error Computation: Multiweight Method Using SPSS-X with Correction for Imputation

```
GET FILE=SYSFILE/               (System file for sample)
      KEEP=DSEX,WEIGHT,SRWT01 TO SRWT56,X.
VECTOR VALUE=MRPCMP1 TO MRPCMP5.
VECTOR WT=SRWT01 TO SRWT56.
VECTOR WX(56).
VECTOR WS(5).
SELECT IF (NOT SYSMIS(MRPCMP1)).
COMPUTE WTX=WEIGHT*MRPCMP1.
LOOP #I=1 TO 56.
   COMPUTE WX(#I) = WT(#I)*MRPCMP1.
END LOOP.
LOOP #I=1 TO 5.
      COMPUTE WS(#I) = VALUE(#I)*WEIGHT.
END LOOP.
AGGREGATE  OUTFILE=*/BREAK=DSEX/UWN=N(WEIGHT)/
      SWT,SW1 TO SW56 = SUM(WEIGHT,SRWT01 TO SRWT56)/
      SWX,SX1 TO SX56 = SUM(WTX,WX1 TO WX56)/
      SS1 TO SS5 = SUM(WS1 TO WS5)/.
VECTOR SW = SW1 TO SW56.
VECTOR SX = SX1 TO SX56.
VECTOR SS = SS1 TO SS5.
COMPUTE XBAR = SWX/SWT.
COMPUTE XVAR = 0.
LOOP #I=1 TO 56.
   COMPUTE #DIFF = SX(#I)/SW(#I) - XBAR.
   COMPUTE XVAR = XVAR + #DIFF * #DIFF.
END LOOP.
LOOP #I=1 TO 5.
      COMPUTE SS(#I) = SS(#I)/SWT.
END LOOP.
COMPUTE SVAR = VARIANCE(SS1 TO SS5).
COMPUTE XSE = SQRT(XVAR+(6/5)*SVAR).
PRINT FORMATS XVAR,SVAR,XSE (F8.4).
LIST VARIABLES=DSEX,UWN,SWT,XBAR,XVAR,SVAR,XSE.
FINISH.
```

Table 10-8

Standard Error Computation: Multiweight Method Using SAS with Correction for Imputation

```
DATA A;
      SET SYSFILE.TSASTUD;
      ARRAY WT SRWT01-SRWT56;
      ARRAY WX WX1-WX56;
      ARRAY VALUE MRPCMP1-MRPCMP5;
      ARRAY WS WS1-WS5;
      IF (MRPCMP1 NE .);
      WTX = WEIGHT*MRPCMP1;
      DO OVER WT;
            WX = WT*MRPCMP1;
      END;
      DO OVER WS;
            WS = VALUE*WEIGHT;
      END;
PROC SUMMARY;
      CLASS DSEX;
      VAR WEIGHT SRWT01-SRWT56 WTX WX1-WX56 WS1-WS5;
      OUTPUT OUT=B    N(WEIGHT)=UWN
            SUM(WEIGHT WTX SRWT01-SRWT56 WX1-WX56 WS1-WS5)=
                SWT SWX SW1-SW56 SX1-SX56 SS1-SS5;
DATA C;
      SET B;
      ARRAY SW SW1-SW56;
      ARRAY SX SX1-SX56;
      ARRAY SS SS1-SS5;
      XBAR = SWX/SWT;
      XVAR = 0;
      DO OVER SW;
            DIFF = (SX/SW)-XBAR;
            XVAR = XVAR+DIFF*DIFF;
      END;
      DO OVER SS;
            SS = SS/SWT;
      END;
      SVAR = VAR(SS1,SS2,SS3,SS4,SS5);
      XSE = SQRT(XVAR+(6/5)*SVAR);
PROC PRINT;
      VAR DSEX UWN SWT XBAR XVAR SVAR XSE;
```

In Tables 10-5 and 10-6, the variable X may be any variable or transformation of variables except plausible values. In Tables 10-7 and 10-8, the vector or array named VALUE refers to a set of plausible values from any of the subject areas.

10.7 AN ANALYSIS EXAMPLE USING 1990 NAEP DATA WITH SAS

In Chapter 1, we explained how to perform an analysis of NAEP data using any statistical or procedural language, and presented an example of how to produce a simple descriptive analysis table that did not include standard error estimates.

This section explains how to use SAS to perform the same analysis, this time including standard error estimates that account for NAEP sampling design and measurement error components. Such an accounting is required for statistical comparison of NAEP data. Because the NAEP sample is not a simple random sample, ordinary formulas for estimating the standard error of sample statistics will produce values that are too small.

Before attempting any analysis of NAEP data, users should understand the special characteristics of the NAEP sampling design (Chapters 2 and 4). Alternate methods for computing standard errors and recommended formulas for obtaining degrees of freedom are given in Chapter 4.

The analysis in our example produced the following estimate, with standard errors, of the reported amount of television watched each day by eighth-grade girls in the national winter public-school sample and the corresponding mean reading proficiency scores. A similar table for each state is included at the beginning of each state's codebook.

Table 10-9
SAS Analysis Example Using Jackknifed Standard Error Estimates

```
                          1990 NATIONAL WINTER PUBLIC SCHOOL SAMPLE
                           MATHEMATICS RESULTS FOR 8TH GRADE GIRLS
                              BY AMOUNT OF TELEVISION VIEWING

       HOW MUCH
       TELEVISION DO
       YOU USUALLY
 OBS   WATCH                  N     WTD  N        PCT    SE(PCT)      MEAN    SE(MEAN)

  1    TOTAL               1403    1425.66    100.000    0.00000   260.462      1.3199
  2    NONE                  11       8.65      0.607    0.18649   259.819     10.1484
  3    1 HOUR OR LESS       184     188.37     13.213    1.05406   269.138      2.8130
  4    2 HOURS              291     284.80     19.977    1.29962   268.956      2.1538
  5    3 HOURS              310     334.46     23.460    1.35049   264.125      1.7599
  6    4 HOURS              233     245.27     17.204    1.46723   260.077      2.3215
  7    5 HOURS              154     153.44     10.763    0.95371   255.362      2.7075
  8    6 HOURS OR MORE      220     210.66     14.776    1.17891   239.590      2.2083
```

BEST COPY AVAILABLE

Most analyses of NAEP data can be performed in four basic steps:

- Identify and access the appropriate data file
- Identify and extract the relevant variables
- Select the proper subset of students
- Compute and print the results

The method you choose to perform these steps may vary with the complexity of the analysis or with the statistical or procedural language you are using.

To begin the example analysis, you need to identify

- the file that contains response data for eighth-grade students and
- the relevant variables in the file.

NAEP files are described in Chapter 9 and listed in Table 9-1; the correct file for our example is NWPSTUD.DAT. Next, find the data set record layout for NWPSTUD.DAT in the accompanying document entitled *Layouts and Codebooks*. Here you will find the names and file locations of the variables needed to produce this table (response counts for each variable are found in the corresponding codebook). To produce the table, we need three variables for the basic data, 56 replicate weight variables to produce the standard errors, and five plausible mathematics values:

Seq. No.	Field Name	Column Position	Field Width	Decimal Places	Type	Range	Short Label
27	DSEX	56	1	--	D	1 - 2	GENDER
47	WEIGHT	110	7	5	C	--	OVERALL STUDENT FULL-SAMPLE WEIGHT
54 • • 109	SRWT01 • • SRWT56	142 • • 527	7 • • 7	5 • • 5	C • • C	--	STUDENT REPLICATE WEIGHT 01 • • STUDENT REPLICATE WEIGHT 56
279	B001801A	1361	1	--	D	1 - 7	HOW MUCH TELEVISION DO YOU USUALLY WATCH EACH DAY
265 • • 269	MRPCMP1 • • MRPCMP5	1327 • • 1347	5 • • 5	2 • • 2	C • • C	--	PLAUSIBLE NAEP MATH VALUE #1 (COMPOSITE) • • PLAUSIBLE NAEP MATH VALUE #5 (COMPOSITE)

For analyses that are relatively simple (requiring the use of just a few variables), you can manually enter the variable labels and locations into your computer program. This example can be performed more efficiently through the use of the SAS control statement files.

To aid users, we have added three types of files to the data tapes:

- machine-readable catalog files
- SAS control statement files
- SPSS-X control statement files

The SAS control statement files are provided to facilitate the creation of SAS system files. There is a SAS control file for each data file on the tapes. Part of each control file contains the field name, location, and format for each variable on the corresponding data file. More about control statement files can be found in Sections 10.2 through 10.4.

Any statistical computing language or package can be used to access this file, extract the relevant variables, select the proper subset of students (grade 8), and compute the values shown in the table. Using SAS, the following procedure will complete the analysis example.

1) Select the file containing the grade 8 students. This is the student main sample described in Table 9-1; its file name is NWPSTUD.DAT. Identify the relevant variables from the data set record layout: DSEX, WEIGHT, SRWT01-SRWT56, B001801A, and MRPCMP1-MRPCMP5.

2) Using the raw data file, select the appropriate subset of students for the table. This selection restricts the analysis to females (DSEX=2) who have valid MRPCMP1 (mathematics proficiency) and B001801A (television viewing) values, and are in grade 8.

3) Compute weighted products and sums corresponding to the 56 student replicate weights and the five estimates of student mathematics proficiency.

4) Compute overall weighted sums for use in the computation of percentages and jackknifed standard errors.

5) Computed weighted sums for each level of television viewing (B001801A).

6) Merge the weighted sums from steps 4 and 5 and compute percentages, variances, and jackknifed standard errors (with sampling and measurement error components).

7) Print the final result in a formatted table.

The SAS code for performing steps 2 to 7 is shown in Table 10-10.

Table 10-10

SAS Code for Steps 2 through 7 to Produce Sample Analysis

```
TITLE1 '1990 NATIONAL WINTER PUBLIC SCHOOL SAMPLE';
TITLE2 'MATHEMATICS RESULTS FOR 8TH GRADE GIRLS';
TITLE3 'BY AMOUNT OF TELEVISION VIEWING';
DATA    A;
INFILE  RAWDATA;
INPUT
  DSEX         56          WEIGHT      110-116 5  SRWT01       142-148 5
  SRWT02       149-155 5   SRWT03      156-162 5  SRWT04       163-169 5
  SRWT05       170-176 5   SRWT06      177-183 5  SRWT07       184-190 5
  SRWT08       191-197 5   SRWT09      198-204 5  SRWT10       205-211 5
  SRWT11       212-218 5   SRWT12      219-225 5  SRWT13       226-232 5
  SRWT14       233-239 5   SRWT15      240-246 5  SRWT16       247-253 5
  SRWT17       254-260 5   SRWT18      261-267 5  SRWT19       268-274 5
  SRWT20       275-281 5   SRWT21      282-288 5  SRWT22       289-295 5
  SRWT23       296-302 5   SRWT24      303-309 5  SRWT25       310-316 5
  SRWT26       317-323 5   SRWT27      324-330 5  SRWT28       331-337 5
  SRWT29       338-344 5   SRWT30      345-351 5  SRWT31       352-358 5
  SRWT32       359-365 5   SRWT33      366-372 5  SRWT34       373-379 5
  SRWT35       380-386 5   SRWT36      387-393 5  SRWT37       394-400 5
  SRWT38       401-407 5   SRWT39      408-414 5  SRWT40       415-421 5
  SRWT41       422-428 5   SRWT42      429-435 5  SRWT43       436-442 5
  SRWT44       443-449 5   SRWT45      450-456 5  SRWT46       457-463 5
  SRWT47       464-470 5   SRWT48      471-477 5  SRWT49       478-484 5
  SRWT50       485-491 5   SRWT51      492-498 5  SRWT52       499-505 5
  SRWT53       506-512 5   SRWT54      513-519 5  SRWT55       520-526 5
  SRWT56       527-533 5
  B001801A     1361        MRPCMP1    1327-1331 2 MRPCMP2     1332-1336 2
  MRPCMP3     1337-1341 2 MRPCMP4    1342-1346 2 MRPCMP5     1347-1351 2;
ARRAY WT      SRWT01-SRWT56;
ARRAY WX      WX1-WX56;
ARRAY VALUE   MRPCMP1-MRPCMP5;
ARRAY WS      WS1-WS5;
IF (MRPCMP1 NE .);
IF (DSEX  EQ 2);
IF (B001801A NE .) AND
   (B001801A GT 0) AND
   (B001801A LT 8);
WTX = WEIGHT*MRPCMP1;
DO OVER WT;
     WX = WT*MRPCMP1;
END;
DO OVER WS;
     WS = VALUE*WEIGHT;
END;
```

(code continued on next page)

Table 10-10 (continued)

SAS Code for Steps 2 through 7 to Produce Sample Analysis

```
MDUMMY = 0;
KEEP  WEIGHT DSEX B001801A SRWT01-SRWT56 MRPCMP1-MRPCMP5
      WX1-WX56 WS1-WS5 WTX MDUMMY;
LABEL
  DSEX     = 'GENDER                                  '
  WEIGHT   = 'OVERALL STUDENT FULL-SAMPLE WEIGHT      '
  SRWT01   = 'STUDENT REPLICATE WEIGHT 01             '
  B001801A = 'HOW MUCH TELEVISION DO YOU USUALLY WATCH'
  MRPCMP1  = 'PLAUSIBLE NAEP MATH VALUE #1 (COMPOSITE)';
PROC FORMAT;
  VALUE DSEX      1='MALE                '    2='FEMALE              ';
  VALUE B001801A  .='TOTAL               '    1='NONE                '
                  2='1 HOUR OR LESS      '    3='2 HOURS             '
                  4='3 HOURS             '    5='4 HOURS             '
                  6='5 HOURS             '    7='6 HOURS OR MORE     ';
PROC SUMMARY;
    VAR MDUMMY WEIGHT SRWT01-SRWT56;
    OUTPUT OUT=B   SUM(MDUMMY)=MDUMMY
        SUM(WEIGHT SRWT01-SRWT56) = TOTSWT TOTSW1-TOTSW56;
PROC SUMMARY DATA=A;
    CLASS B001801A;
    VAR WEIGHT SRWT01-SRWT56
        WTX WX1-WX56 WS1-WS5
        MDUMMY;
    OUTPUT OUT=C    N(WEIGHT)=UWN
        N(SRWT01-SRWT56) = NSW1-NSW56
        SUM(WEIGHT SRWT01-SRWT56 WTX WX1-WX56 WS1-WS5) =
            SWT    SW1-SW56      SWX SX1-SX56 SS1-SS5
        SUM(MDUMMY) = MDUMMY;
DATA D;
    MERGE B C;
    BY MDUMMY;
    ARRAY SW     SW1-SW56;
    ARRAY TOTSW  TOTSW1-TOTSW56;
    ARRAY SX     SX1-SX56;
    ARRAY SS     SS1-SS5;
    P = 100.0*SWT/TOTSWT;
    XBAR = SWX/SWT;
    XVAR = 0;
    DO OVER SW;
         DIFF = (SX/SW)-XBAR;
         XVAR = XVAR+DIFF*DIFF;
    END;
```

(code continued on next page)

Table 10-10 (continued)

SAS Code for Steps 2 through 7 to Produce Sample Analysis

```
    DO OVER SS;
         SS = SS/SWT;
    END;
    SVAR = VAR(SS1,SS2,SS3,SS4,SS5);
    XSE = SQRT(XVAR+(6/5)*SVAR);

    PSUM = 0;
    DO OVER SW;
         DIFF = 100.0*(SW/TOTSW)-P;
         PSUM = PSUM+DIFF*DIFF;
    END;
    SE = SQRT(PSUM);
PROC PRINT SPLIT='*';
    FORMAT B001801A B001801A.;
    LABEL UWN = 'N'
          SWT = 'WTD  N'
          P   = 'PCT'
          SE  = 'SE(PCT)'
          XBAR= 'MEAN'
          XSE = 'SE(MEAN)';
    VAR  B001801A UWN SWT P SE XBAR XSE;
```

Appendix A

NAEP HISTORY

APPENDIX A

NAEP HISTORY

The National Assessment of Educational Progress (NAEP) is a continuing, congressionally mandated national survey of the knowledge, skills, understandings, and attitudes of young Americans in major subject areas usually taught in school. Its primary goals are to detect and report the current status of, as well as changes in, the educational attainments of young Americans, and to report long-term trends in those attainments. The purpose of NAEP is to gather information that will aid educators, legislators, and others in improving the educational experience of youth in the United States. It is the first ongoing effort to obtain comprehensive and dependable achievement data on a national basis in a uniform, scientific manner.

Between 1964 and 1969, initial assessment planning and development activities were conducted for NAEP with support from both the Carnegie Corporation and the Ford Foundation. During this time, objectives and exercises were developed for many of the subject areas, sampling and data collection strategies were planned, and data analysis plans were formulated and outlined.

From its inception, NAEP has developed assessments through a consensus process. Educators, scholars, and laypersons design objectives for each subject area, proposing general goals they think Americans should achieve in the course of their education. After careful reviews, the objectives are given to item writers, who develop measurement instruments appropriate to the objectives.

After the items pass extensive reviews by subject matter specialists, measurement experts, and laypersons and are pretested in a sample of schools throughout the country, they are administered to a stratified multistage national probability sample. The young people sampled are selected so that assessment results may be generalized to the entire national population.

NAEP collected data for the first time in 1969. Since that time, samples have included over one million 9-, 13- and 17-year-old students and, as funding would allow, 17-year-olds who had left school and adults 26 to 35 years of age. In 1984, grade samples of students were added to the assessment. As Table A-1 illustrates, assessments have focused on traditional subject areas such as reading, writing, mathematics, science, and U.S. history and on less traditional areas such as citizenship, art, literature, music, computer competence, and career and occupational development.

Since 1971, NAEP has been solely supported by federal funds. Funding agencies have included the Office of Education, the National Center for Education, and the National Institute of Education. NAEP is currently supported by the U.S. Department of Education's Office of Educational Research and Improvement, National Center for Education Statistics.

NAEP was administered by the Education Commission of the States (ECS) through 1982. In 1983, Educational Testing Service (ETS) assumed responsibility for administration of the project, incorporating an updated sampling design and, at the same time, making a concerted effort to provide continuity with previous assessments.

Secondary-use data files were first produced in 1975, allowing outside researchers access to the NAEP database. In June 1985, ETS produced its first public-use data files, in a new format, for the 1984 assessment. The new format produced by ETS makes the tapes easier to use (e.g., files have been more simply organized, documentation has been improved and made more accessible).

Table A-1

National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969-1990

Assessment Year	Subject Area(s)	Grades/Ages Assessed										
		Grade 3	Grade 4	Age 9	Grade 7	Grade 8	Age 13	Grade 11	Grade 12	Age 17	Age 17 OS[a]	Adult
1969-70	Science			X			X			X	X	X
1970-71	Reading			X			X			X	X	X
	Literature			X			X			X	X	X
1971-72	Music			X			X			X	X	X
	Social Studies			X			X			X	X	X
1972-73	Science			X			X			X	X	X
	Mathematics			X			X			X	X	X
1973-74	Career and Occupational Development			X			X			X	X	X
	Writing			X			X			X	X	
1974-75	Reading			X			X			X	X	
	Art			X			X			X	X	
	Index of Basic Skills									X	X	
1975-76	Citizenship/Social Studies			X			X			X	X	
	Mathematics[b]						X			X	X	
1976-77	Science			X			X			X		
	Basic Life Skills[b]									X		
	Science, Reading, Health[b]											X
1977-78	Mathematics			X			X			X		
	Consumer Skills[b]									X		
1978-79	Writing, Art, and Music			X			X			X		
1979-80[c]	Reading/Literature			X			X			X	X	
	Art						X					
1981-82	Science[b]			X			X			X		
	Mathematics and Citizenship/Social Studies			X			X			X		

Table A-1 (continued)

National Assessment of Educational Progress
Subject Areas, Grades, and Ages Assessed: 1969-1990

Assessment Year	Subject Area(s)	Grades/Ages Assessed										
		Grade 3	Grade 4	Age 9	Grade 7	Grade 8	Age 13	Grade 11	Grade 12	Age 17	Age 17 OS[a]	Adult
1984[c]	Reading		X	X		X	X	X		X		
	Writing		X	X		X	X	X		X		
1985[c]	Adult Literacy[b]											X
1986[c]	Reading	X		X	X		X	X		X		
	Mathematics	X		X	X		X	X		X		
	Science	X		X	X		X	X		X		
	Computer Competence	X		X	X		X	X		X		
	U.S. History[b]							X		X		
	Literature[b]							X		X		
1988[c]	Reading		X	X		X	X		X	X		
	Writing		X	X		X	X		X	X		
	Civics		X	X		X	X		X	X		
	U.S. History		X	X		X	X		X	X		
	Document Literacy[b]					X	X		X	X		
	Geography[b]								X	X		
	Mathematics[b]			X			X	X		X		
	Science[b]			X			X	X		X		
1990[c]	Reading		X	X		X	X		X	X		
	Mathematics		X	X		X	X		X	X		
	Science		X	X		X	X		X	X		
	Writing[b]		X	X		X	X	X		X		

---

[a] Age 17 students who had dropped out of school or had graduated prior to assessment.
[b] Small, special-interest assessment conducted on limited samples at specific grades or ages.
[c] Assessment conducted by Educational Testing Service.

BEST COPY AVAILABLE

Appendix B

1990 TRIAL STATE ASSESSMENT IRT PARAMETERS

APPENDIX B

1990 TRIAL STATE ASSESSMENT IRT PARAMETERS

This appendix contains five tables of IRT (item response theory) parameters for NAEP items that were used in the mathematics subscales for the 1990 Trial State Assessment.

For each NAEP item used in scaling, the tables show the corresponding IRT parameters (A, B, and C) and standard errors (S.E.), the block in which the item appears for each age class (BLOCK), and the order in which the item appears within the block (ITEM).

Note that item parameters shown in this appendix are in the metrics used for the original calibration of the scale. The transformations needed to represent these parameters in terms of the metric of the final reporting scales are given in Chapter 10 of *The Technical Report of NAEP's 1990 Trial State Assessment*.

Table B-1

IRT Parameters for Mathematics Items:
Numbers and Operations

NAEP ID	A (S.E.)	B (S.E.)	C (S.E.)	Block	Item
M011131	0.643 (0.022)	-1.477 (0.098)	0.155 (0.042)	M8	13
M012431	0.828 (0.024)	-0.396 (0.040)	0.080 (0.019)	M8	3
M012531	0.661 (0.025)	0.655 (0.033)	0.066 (0.013)	M8	4
M012931	0.919 (0.050)	1.213 (0.026)	0.212 (0.009)	M8	8
M013431	0.956 (0.037)	0.191 (0.032)	0.131 (0.014)	M8	15
M013531	0.638 (0.044)	1.796 (0.045)	0.085 (0.010)	M8	16
M013631	1.344 (0.052)	0.937 (0.015)	0.058 (0.005)	M8	17
M015501	0.969 (0.033)	0.224 (0.026)	0.082 (0.012)	M7	2
M015901	0.685 (0.047)	1.246 (0.039)	0.219 (0.014)	M7	6
M016501	1.075 (0.061)	1.695 (0.030)	0.079 (0.005)	M7	12
M017401	0.258 (0.016)	-5.220 (0.387)	0.198 (0.057)	M4	1
M017701	0.844 (0.025)	-1.050 (0.057)	0.125 (0.029)	M4	4
M017901	1.147 (0.035)	-0.892 (0.038)	0.105 (0.023)	M4	6
M018201	0.601 (0.018)	-0.756 (0.064)	0.090 (0.025)	M4	9
M018401	1.202 (0.050)	-0.743 (0.050)	0.322 (0.024)	M4	11
M018501	1.620 (0.067)	0.541 (0.017)	0.237 (0.007)	M4	12
M018601	0.598 (0.036)	1.201 (0.040)	0.135 (0.015)	M4	13
M020001	0.667 (0.013)	-0.214 (0.014)	0.000 (0.000)	M5	4
M020101	1.304 (0.025)	-0.329 (0.009)	0.000 (0.000)	M5	5
M020501	0.847 (0.016)	-0.390 (0.012)	0.000 (0.000)	M5	9
M021901	0.868 (0.025)	-1.387 (0.063)	0.135 (0.035)	M6	1
M022001	1.025 (0.030)	-0.802 (0.043)	0.135 (0.024)	M6	2
M022301	0.626 (0.021)	-2.456 (0.113)	0.176 (0.051)	M6	5
M022701	0.859 (0.029)	-0.813 (0.060)	0.170 (0.029)	M6	9
M022901	1.161 (0.046)	-0.292 (0.039)	0.312 (0.017)	M6	12
M023001	1.105 (0.039)	-0.230 (0.034)	0.225 (0.016)	M6	13
M023801	1.261 (0.043)	0.310 (0.019)	0.101 (0.009)	M6	21
M027031	0.402 (0.023)	-4.564 (0.263)	0.193 (0.056)	M9	1
M027331	0.778 (0.014)	0.600 (0.014)	0.000 (0.000)	M9	4
M027831	1.013 (0.017)	0.059 (0.010)	0.000 (0.000)	M9	9
M028031	0.950 (0.040)	0.622 (0.027)	0.181 (0.011)	M9	11
M028131	0.541 (0.012)	0.832 (0.021)	0.000 (0.000)	M9	12
M028231	0.687 (0.025)	0.486 (0.035)	0.060 (0.014)	M9	13
M028631	1.276 (0.033)	1.499 (0.018)	0.000 (0.000)	M9	17
M028731	1.729 (0.103)	1.541 (0.020)	0.082 (0.003)	M9	18
M028931	0.629 (0.058)	1.430 (0.058)	0.258 (0.018)	M9	20
N202831	0.627 (0.023)	-1.998 (0.125)	0.198 (0.054)	M8	12
N258801	1.167 (0.068)	0.668 (0.031)	0.411 (0.010)	M3	11

Table B-1 (continued)

IRT Parameters for Mathematics Items:
Numbers and Operations

NAEP ID	A (S.E.)	B (S.E.)	C (S.E.)	Block	Item
N260101	1.075 (0.041)	-0.245 (0.039)	0.228 (0.018)	M3	18
N274801	0.685 (0.041)	-0.284 (0.105)	0.417 (0.029)	M3	10
N275301	0.280 (0.014)	-3.068 (0.263)	0.172 (0.053)	M3	14
N276803	0.223 (0.011)	-3.735 (0.176)	0.000 (0.000)	M3	1
N277602	0.418 (0.012)	-2.415 (0.065)	0.000 (0.000)	M3	2
N286201	0.806 (0.027)	-1.157 (0.074)	0.151 (0.037)	M3	6
N286301	1.112 (0.042)	0.178 (0.029)	0.180 (0.013)	M3	21
N286602	0.641 (0.012)	-0.141 (0.014)	0.000 (0.000)	M3	13

Table B-2

IRT Parameters for Mathematics Items:
Measurement

NAEP ID	A (S.E.)	B (S.E.)	C (S.E.)	Block	Item
M012331	0.717 (0.035)	-1.427 (0.116)	0.200 (0.051)	M8	2
M013331	0.878 (0.048)	-1.356 (0.105)	0.211 (0.052)	M8	14
M015401	0.710 (0.051)	0.043 (0.085)	0.190 (0.032)	M7	1
M015701	0.837 (0.039)	-2.000 (0.111)	0.227 (0.058)	M7	4
M016201	0.887 (0.077)	0.787 (0.041)	0.211 (0.017)	M7	9
M017501	0.431 (0.025)	-2.430 (0.232)	0.288 (0.063)	M4	2
M018101	0.804 (0.062)	-0.073 (0.090)	0.269 (0.033)	M4	8
M019101	1.482 (0.241)	2.032 (0.072)	0.175 (0.006)	M4	18
M019201	1.450 (0.205)	1.894 (0.061)	0.147 (0.006)	M4	19
M020301	1.000 (0.030)	-0.354 (0.014)	0.000 (0.000)	M5	7
M022601	1.129 (0.108)	0.780 (0.037)	0.381 (0.013)	M6	8
M022801	1.751 (0.057)	-0.608 (0.012)	0.000 (0.000)	M6	10
M022802	1.604 (0.048)	-0.929 (0.015)	0.000 (0.000)	M6	11
M023401	0.860 (0.075)	0.295 (0.069)	0.364 (0.024)	M6	17
M023701	0.519 (0.017)	1.038 (0.033)	0.000 (0.000)	M6	20
M027631	1.067 (0.086)	0.094 (0.053)	0.209 (0.024)	M9	7
M028831	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	M9	19
N252101	0.654 (0.062)	0.275 (0.107)	0.268 (0.035)	M3	17
N265201	0.755 (0.044)	-1.872 (0.158)	0.339 (0.066)	M3	9
N265901	0.742 (0.069)	0.651 (0.066)	0.250 (0.024)	M3	16
N267201	0.796 (0.061)	-1.009 (0.162)	0.401 (0.056)	M3	3

Table B-3

IRT Parameters for Mathematics Items:
Geometry

NAEP ID	A (S.E.)	B (S.E.)	C (S.E.)	Block	Item
M012731	0.646 (0.058)	1.325 (0.053)	0.174 (0.019)	M8	6
M012831	1.185 (0.071)	0.649 (0.026)	0.119 (0.012)	M8	7
M015601	0.358 (0.030)	-0.078 (0.251)	0.234 (0.054)	M7	3
M016301	0.608 (0.028)	-0.289 (0.085)	0.123 (0.032)	M7	10
M016401	1.580 (0.118)	1.234 (0.022)	0.167 (0.006)	M7	11
M016601	0.833 (0.049)	1.375 (0.031)	0.080 (0.010)	M7	13
M016701	1.236 (0.093)	1.718 (0.034)	0.119 (0.005)	M7	14
M017601	0.459 (0.022)	-1.744 (0.169)	0.184 (0.054)	M4	3
M018001	0.755 (0.049)	0.044 (0.084)	0.218 (0.031)	M4	7
M019001	0.733 (0.050)	0.776 (0.050)	0.150 (0.020)	M4	17
M019601	0.720 (0.063)	1.650 (0.047)	0.128 (0.013)	M4	21
M019801	0.982 (0.023)	-0.578 (0.014)	0.000 (0.000)	M5	2
M019901	0.675 (0.018)	-1.438 (0.032)	0.000 (0.000)	M5	3
M020901	0.563 (0.016)	1.314 (0.033)	0.000 (0.000)	M5	11
M021001	0.862 (0.019)	0.277 (0.013)	0.000 (0.000)	M5	12
M021301	1.194 (0.027)	0.125 (0.011)	0.000 (0.000)	M5	15
M021302	1.165 (0.026)	-0.079 (0.011)	0.000 (0.000)	M5	16
M022201	0.539 (0.015)	-0.645 (0.023)	0.000 (0.000)	M6	4
M022501	0.800 (0.020)	-0.368 (0.015)	0.000 (0.000)	M6	7
M023101	1.087 (0.053)	0.029 (0.038)	0.118 (0.019)	M6	14
M027231	0.704 (0.057)	-0.303 (0.144)	0.401 (0.042)	M9	3
M027431	0.669 (0.033)	-0.627 (0.102)	0.167 (0.040)	M9	5
M028331	1.595 (0.228)	1.602 (0.042)	0.351 (0.007)	M9	14
N253701	0.525 (0.042)	-0.309 (0.194)	0.309 (0.052)	M3	12
N254602	1.322 (0.100)	1.029 (0.024)	0.196 (0.009)	M3	22
N269901	0.816 (0.061)	-0.152 (0.104)	0.337 (0.036)	M3	15

Table B-4

IRT Parameters for Mathematics Items: Data Analysis, Statistics, and Probability

NAEP ID	A (S.E.)	B (S.E.)	C (S.E.)	Block	Item
M012631	1.983 (0.153)	0.788 (0.017)	0.216 (0.008)	M8	5
M013031	1.167 (0.041)	1.508 (0.029)	0.000 (0.000)	M8	9
M013131	0.952 (0.032)	1.390 (0.029)	0.000 (0.000)	M8	10
M015801	1.074 (0.060)	0.436 (0.031)	0.116 (0.015)	M7	5
M016101	1.429 (0.094)	0.481 (0.027)	0.246 (0.012)	M7	8
M017001	0.860 (0.070)	1.183 (0.032)	0.140 (0.013)	M7	18
M017801	1.198 (0.084)	-0.228 (0.064)	0.304 (0.026)	M4	5
M018901	1.207 (0.224)	2.063 (0.138)	0.157 (0.007)	M4	16
M020201	0.576 (0.019)	-2.059 (0.056)	0.000 (0.000)	M5	6
M020801	1.140 (0.048)	1.630 (0.037)	0.000 (0.000)	M5	10
M021101	0.944 (0.025)	0.157 (0.013)	0.000 (0.000)	M5	13
M023301	1.792 (0.120)	-0.459 (0.045)	0.247 (0.023)	M6	16
M023501	1.920 (0.142)	0.834 (0.015)	0.123 (0.007)	M6	18
M023601	0.895 (0.040)	-0.366 (0.055)	0.093 (0.025)	M6	19
M028531	0.981 (0.029)	-0.777 (0.022)	0.000 (0.000)	M9	16
N250201	0.668 (0.031)	-1.437 (0.124)	0.175 (0.051)	M3	8
N250901	0.333 (0.018)	-3.623 (0.256)	0.175 (0.054)	M3	4
N250902	0.829 (0.033)	-0.881 (0.069)	0.104 (0.032)	M3	5
N263501	1.368 (0.082)	0.104 (0.035)	0.214 (0.016)	M3	19

Table B-5

IRT Parameters for Mathematics Items: Algebra and Functions

NAEP ID	A (S.E.)	B (S.E.)	C (S.E.)	Block	Item
M012231	0.436 (0.027)	-3.985 (0.236)	0.148 (0.051)	M8	1
M013231	1.180 (0.116)	1.916 (0.055)	0.123 (0.006)	M8	11
M013731	0.925 (0.079)	1.520 (0.042)	0.117 (0.010)	M8	18
M016001	0.919 (0.038)	0.475 (0.029)	0.065 (0.013)	M7	7
M016801	0.949 (0.053)	1.766 (0.038)	0.040 (0.005)	M7	15
M016901	2.279 (0.000)	0.862 (0.012)	0.161 (0.000)	M7	16
M016902	1.719 (0.000)	1.170 (0.011)	0.000 (0.000)	M7	17
M018301	0.842 (0.035)	-0.411 (0.062)	0.132 (0.028)	M4	10
M018701	1.334 (0.073)	0.318 (0.030)	0.223 (0.013)	M4	14
M018801	0.840 (0.071)	1.122 (0.041)	0.277 (0.015)	M4	15
M019301	1.192 (0.089)	1.300 (0.028)	0.191 (0.008)	M4	20
M019701	0.510 (0.016)	-1.641 (0.046)	0.000 (0.000)	M5	1
M020401	0.637 (0.016)	0.029 (0.016)	0.000 (0.000)	M5	8
M021201	1.020 (0.026)	0.599 (0.013)	0.000 (0.000)	M5	14
M022101	0.739 (0.035)	-2.689 (0.126)	0.222 (0.058)	M6	3
M022401	1.098 (0.069)	-0.575 (0.082)	0.391 (0.033)	M6	6
M023201	0.998 (0.042)	-0.435 (0.050)	0.124 (0.025)	M6	15
M027131	0.843 (0.030)	-1.989 (0.076)	0.124 (0.043)	M9	2
M027531	0.627 (0.033)	-0.803 (0.126)	0.221 (0.045)	M9	6
M027731	0.864 (0.041)	0.197 (0.044)	0.117 (0.019)	M9	8
M027931	0.977 (0.022)	0.093 (0.012)	0.000 (0.000)	M9	10
M028431	0.721 (0.019)	0.786 (0.020)	0.000 (0.000)	M9	15
N255701	1.227 (0.070)	0.749 (0.023)	0.132 (0.011)	M3	23
N256101	0.925 (0.025)	-1.189 (0.023)	0.000 (0.000)	M3	7
N264701	1.544 (0.091)	0.481 (0.024)	0.186 (0.012)	M3	20

Appendix C

GLOSSARY OF TERMS

APPENDIX C
Glossary of Terms

anchoring. The process of characterizing score levels in terms of predicted observable behavior.

assessment session. The period of time during which a NAEP booklet is administered to one or more individuals.

background questionnaires. The instruments used to collect information about students' demographics and educational experiences.

bias. In statistics, the difference between the expected value of an estimator and the population parameter being estimated. If the average value of the estimator over all possible samples (the estimator's expected value) equals the parameter being estimated, the estimator is said to be **unbiased**; otherwise, the estimator is **biased**.

BIB (Balanced Incomplete Block) spiraling. A complex variant of multiple matrix sampling, in which items are administered in such a way that each pair of items is administered to a nationally representative sample of respondents.

BILOG. A computer program for estimating item parameters.

block. A group of assessment items created by dividing the item pool for an age/grade into subsets. Used in the implementation of the BIB spiral sample design.

booklet. The assessment instrument created by combining blocks of assessment items.

calibrate. To estimate the parameters of a set of items from responses of a sample of examinees.

clustering. The process of forming sampling units as groups of other units.

codebook. A formatted printout of NAEP data for a particular sample of respondents.

coefficient of variation. The ratio of the standard deviation of an estimate to the value of the estimate.

common block. A group of background items included in the beginning of every assessment booklet.

conditional probability. Probability of an event, given the occurrence of another event.

conditioning variables. Demographic and other background variables characterizing a respondent. Used in construction of plausible values.

degrees of freedom. [of a variance estimator] The number of independent pieces of information used to generate a variance estimate.

derived variables. Subgroup data that were not obtained directly from assessment responses, but through procedures of interpretation, classification, or calculation.

design effects. The ratio of the variance for the sample design to the variance for a simple random sample of the same size.

distractor. An incorrect response choice included in a multiple-choice item.

excluded student questionnaire. An instrument completed for every student who was sampled but excluded from the assessment.

excluded students. Sampled students determined by the school to be unable to participate because they have limited English proficiency, are mildly mentally retarded (educable), or are functionally disabled.

expected value. The average of the sample estimates given by an estimator over all possible samples. If the estimator is unbiased, then its expected value will equal the population value being estimated.

field test. A pretest of items to obtain information regarding clarity, difficulty levels, timing, feasibility, and special administrative situations; performed before revising and selecting items to be used in the assessment.

focused-BIB spiraling. A variation of BIB spiraling in which items are administered in such a way that each pair of items *within a subject area* is administered to a nationally representative sample of respondents.

foils. The correct and incorrect response choices included in a multiple-choice item.

group effect. The difference between the mean for a group and the mean for the nation.

imputation. Prediction of a missing value according to some procedure, using a mathematical model in combination with available information. See **plausible values**.

imputed race/ethnicity. The race or ethnicity of an assessed student, as derived from his or her responses to particular common background items. A NAEP **reporting subgroup**.

item response theory (IRT). Test analysis procedures that assume a mathematical model for the probability that a given examinee will respond correctly to a given exercise.

jackknife. A procedure to estimate standard errors of percentages and other statistics. Particularly suited to complex sample designs.

machine-readable catalog. Computer processing control information, IRT parameters, foil codes, and labels in a computer-readable format.

major strata. Used to stratify the primary sampling frame within each region. Involves stratification by size of community and degree of ruralization.

metropolitan statistical area (MSA). An area defined by the federal government for the purposes of presenting general-purpose statistics for metropolitan areas. Typically, an MSA contains a city with a population of at least 50,000 plus adjacent areas.

multistage sample design. Indicates more than one stage of sampling. An example of three-stage sampling: 1) sample of counties (primary sampling units or PSUs); 2) sample of schools within each sample county; 3) sample of students within each sample school.

multiple matrix sampling. Sampling plan in which different samples of respondents take different samples of items.

NAEP scales. The anchored scales common across age/grade levels and assessment years used to report NAEP results.

nonresponse. The failure to obtain responses or measurements for all sample elements.

nonsampling error. A general term applying to all sources of error except sampling error. Includes errors from defects in the sampling frame, response or measurement error, and mistakes in processing the data.

objective. A desirable education goal agreed upon by scholars in the field, educators, and concerned laypersons, and established through the consensus approach.

observed race/ethnicity. Race or ethnicity of an assessed student as perceived by the exercise administrator.

open-ended response item. A nonmultiple-choice item that requires some type of written or oral response.

oversampling. Deliberately sampling a portion of the population at a higher rate than the remainder of the population.

parental education. The level of education of the mother and father of an assessed student as derived from the student's response to two assessment items. A NAEP **reporting subgroup**.

percent correct. The percent of a target population that would answer a particular exercise correctly.

plausible values. Proficiency values drawn at random from a conditional distribution of a NAEP respondent, given his or her response to cognitive exercises and a specified subset of background variables (conditioning variables). The selection of a plausible value is a form of **imputation**.

poststratification. Classification and weighting to correspond to external values of selected sampling units by a set of strata definitions after the sample has been selected.

primary sampling unit (PSU). The basic geographic sampling unit for NAEP. Either a single county or a set of contiguous counties.

Principal's Questionnaire. A questionnaire sent to every sampled school that agreed to participate in the Trial State Assessment. It requested aggregate information on enrollment by grade, race, and ethnicity of the student population, community size, and the distribution of employment status of parents of attending students.

probability sample. A sample in which every element of the population has a known, nonzero probability of being selected.

pseudoreplicate. The value of a statistic based on an altered sample. Used by the **jackknife** variance estimator.

QED. Quality Education Data, Inc. A supplier of lists of schools, school districts, and other school data.

random variable. A variable that takes on any value of a specified set with a particular probability.

region. One of four geographic areas used in gathering and reporting data:

Northeast, Southeast, Central, and West (as defined by the Office of Business Economics, U.S. Department of Commerce). A NAEP **reporting subgroup.**

reporting subgroup. Groups within the national population for which NAEP data are reported: for example, gender, race/ethnicity, grade, age, level of parental education, region, and size and type of community.

respondent. A person who is eligible for NAEP, is in the sample, and responds by completing one or more items in an assessment booklet.

response options. In a multiple-choice question, alternatives that can be selected by a respondent.

sample. A portion of a population, or a subset from a set of units, selected by some probability mechanism for the purpose of investigating the properties of the population. NAEP does not assess an entire population but rather selects a representative sample from the group to answer assessment items.

sampling error. The error in survey estimates that occurs because only a sample of the population is observed. Measured by sampling **standard error**.

sampling frame. The list of sampling units from which the sample is selected.

sampling weight. A multiplicative factor equal to the reciprocal of the probability of a respondent being selected for assessment with adjustment for nonresponse and perhaps also for poststratification. The sum of the weights provides an estimate of the number of persons in the population represented by a respondent in the sample.

school characteristics and policy questionnaire. A questionnaire completed for each school by the principal or other official; used to gather information concerning school administration, staffing patterns, curriculum, and student services.

secondary-use data files. Computer files containing respondent-level cognitive, demographic, and background data. Available for use by researchers wishing to perform analyses of NAEP data.

selection probability. The chance that a particular sampling unit has of being selected in the sample.

session. A group of students reporting for the administration of an assessment. Most schools conducted only one session, but some large schools conducted as many as 10 or more.

simple random sample. Process for selecting n sampling units from a population of N sampling units so that each sampling unit has an equal chance of being in the sample and every combination of n sampling units has the same chance of being in the sample chosen.

standard error. A measure of sampling variability and measurement error for a statistic. Because of NAEP's complex sample design, sampling standard errors are estimated by **jackknifing** the samples from first-stage sample estimates. Standard errors may also include a component due to the error of measurement of individual scores estimated using plausible values.

stratification. The division of a population into parts, called strata.

stratified sample. A sample selected from a population that has been stratified, with a sample selected independently in each stratum. The strata are defined for the purpose of reducing sampling error.

student ID number. A unique identification number assigned to each respondent to preserve his or her anonymity. NAEP does not record the names of any respondents.

subject area. One of the areas assessed by National Assessment; for example, art, civics, computer competence, geography, literature, mathematics, music, reading, science, U.S. history, or writing.

systematic sample (systematic random sample). A sample selected by a systematic method; for example, when units are selected from a list at equally spaced intervals.

teacher questionnaire. A questionnaire completed by selected teachers of sample students; used to gather information concerning years of teachin experience, frequency of assignments, teaching materials used, and availability and use of computers.

Trial State Assessment Program. The NAEP program, authorized by Congress in 1988, which was established to provide for a program of voluntary state-by-state assessments on a trial basis.

trimming A process by which extreme weights are reduced (trimmed) to diminish the effect of extreme values on estimates and estimated variances.

type 1 Cluster. Individual schools in states where all eighth-grade schools were included in the sample.

type 2 Cluster. A school or group of schools in states where small schools were grouped geographically with large ones.

type 3A Cluster. A large school in states where large and small schools were stratified and sampled separately.

type 3B Cluster. A group of small schools in states where large and small schools were stratified and sampled separately.

type of community (TOC). One of the NAEP **reporting subgroups**, dividing the communities in the nation into four groups on the basis of the proportion of the students living in each of three sizes of communities and on the percentage of parents in each of five occupational categories.

variance. The average of the squared deviations of a random variable from the expected value of the variable. The variance of an estimate is the squared standard error of the estimate.

REFERENCES CITED IN TEXT

REFERENCES CITED IN TEXT

Beaton, A. E. & Johnson, E. G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15*, 9-38.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

Harris, R. J. (1975). *A primer of multivariate statistics*. New York, NJ: Academic Press.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Johnson, E. G., & Rust, K. F. (in press). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, Series B, *36*, 1-22.

Koffler, S. L. (1991). *The technical report of NAEP's 1990 Trial State Assessment*. (No. 21-ST-01) Princeton, NJ: Educational Testing Service; Washington, DC: National Center for Education Statistics.

Little, R. J. A. & Rubin, D. B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37*, 218-220.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Miller, R. G. (1981). *Simultaneous statistical inference* (2nd ed.). New York: Springer.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*, 993-997.

Mislevy, R. J. (1988). *Randomization-based inferences about latent variables from complex samples*. (ETS Research Report RR-88-54-ONR) Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (in press). Randomization-based inference about latent variables from complex samples. *Psychometrika*.

Mislevy, R. J. & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J. & Sheehan, K. M. (1987). Marginal estimation procedures. In A.E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report* (No 15-TA-20) Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Mislevy, R. J. & Wu, P-K. (1988). *Inferring examinee ability when some item responses are missing.* (ETS Research Report RR-88-48-ONR) Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Satterthwaite, F. E. (1946). An approximate distribution for estimates of variance components. *Biometrics*, 2, 110-114.

Sheehan, K. M. (1985). *M-GROUP: Estimation of group effects in multivariate models* [Computer program] Princeton, NJ: Educational Testing Service.

Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.). (1989). *Analysis of complex surveys.* New York: John Wiley & Sons.

Wingersky, M., Kaplan, B. A. & Beaton, A. E. (1987). Joint estimation procedures. In A.E. Beaton, *Implementing the new design: The NAEP 1983-84 technical report.* (No 15-TR-20) Princeton, NJ: National Association of Educational Progress, Educational Testing Service.